Manuel Lameiras de Figueiredo Campagnolo

PROPOSTA DE UM MÉTODO PARA INTEGRAÇÃO DE CONHECIMENTO EM CLASSIFICAÇÃO

Lisboa, Setembro de 1992

Dissertação apresentada como requisito parcial para obtenção de grau de Mestre em Matemática Aplicada à Economia e à Gestão no Instituto Superior de Economia e Gestão da Universidade Técnica de Lisboa.

O autor foi bolseiro do Instituto Nacional de Investigação Científica durante o período inicial de preparação da dissertação

Agradecimentos

Ao Prof. Helder Coelho pelas críticas pertinentes e pelas orientações sempre preciosas que me ajudaram a estabelecer o enquadramento desta tese e a definir, ao longo do meu trabalho, a direcção de investigação a seguir.

Ao Prof. José Miguel Cardoso Pereira pelo incentivo que me deu para o prosseguimento deste trabalho e pelos dados que me disponibilizou gentilmente.

Aos meus colegas da Secção Autónoma de Matemática do Instituto Superior de Agronomia, e em particular a José Casquilho, Fernanda Valente, Isabel Faria, Manuela Neves e Jorge Orestes, pelo apoio constante, a ajuda, as críticas e o bom ambiente de trabalho e de camaradagem que me proporcionaram. Ao Prof. St. Aubyn a confiança que depositou em mim desde a altura em que me licenciei.

Aos meus pais, que me ensinaram no dia a dia o que é a curiosidade científica e que me mostraram também que há lugar para a generosidade e bondade em todas as coisas, devo grande parte da minha formação. Obrigado também pelos vossos conselhos e pelas vossas críticas.

Ao meu irmão, aos meus avós, às pessoas próximas que me fizeram sentir acarinhado desde sempre.

À Paula, que me apoiou nos momentos difíceis e que encheu a minha vida de felicidade. Obrigado pelas tuas sugestões, pelas tuas critícas, pela tua amizade e ... pela tua participação directa na composição deste texto.

RESUMO

Os problemas considerados são problemas de classificação caracterizados por: 1) os objectos a classificar pertencem a uma de um conjunto de classes prédefinidas; 2) é conhecida a classificação de um conjunto de objectos (amostra); 3) todos os objectos são descritos por um conjunto de atributos. Esta categoria de problemas designa-se por *problemas de classificação supervisada.*

Neste trabalho é proposto um método que permite tomar em consideração a informação incluida na amostra e conhecimento pericial sobre o domínio dos objectos a classificar.

Os objectivos são: 1) definir uma regra de decisão para classificar a totalidade dos objectos; 2) tirar partido do conhecimento pericial para definir essa regra; 3) captar, em estruturas que descrevem o domínio dos objectos, o conhecimento envolvido no processo de classificação.

A abordagem seguida consiste em desenvolver um método de classificação cujo suporte estrutural é uma árvore de decisão. O método auxilia, no decorrer do processo de classificação, o perito/decisor a formular hipóteses sobre o domínio dos objectos e permite estruturar e integrar o conhecimento pericial e a informação da amostra. O conhecimento resultante é incorporado em duas estruturas, uma árvore de classificação e um grafo bayesiano. O grafo bayesiano é um modelo qualitativo (gráfico) e quantitativo (probabilístico) de relações de causa-efeito entre as variáveis envolvidas.

É apresentada uma implementação do método em C-PROLOG e uma aplicação a um problema real.

Índice

1 . Introdução

1.1. Motivações

Considere-se o seguinte problema hipotético. Defina-se uma determinada área geográfica e divida-se essa área em parcelas de terreno não sobrepostas. Suponha-se que são conhecidas determinadas características de todas as parcelas como, por exemplo, a altitude, o declive e a reflectância para diversas bandas de comprimento de onda (obtidas, por exemplo, por detecção remota). Definam-se as seguintes classes de ocupação do solo: urbano, agricultura e floresta. Suponha-se também que é conhecida, para uma amostra de parcelas de terreno, a verdadeira classe. O problema consiste em afectar cada uma das parcelas a uma determinada classe, isto é, para cada parcela, para a qual é conhecida a altitude, o declive e as reflectâncias, prevêr o respectivo tipo de ocupação do solo.

O problema acima descrito é um problema de **classificação**. Existem diversas formas de resolver esse problema. Considerem-se, em seguida, duas abordagens frequentes mas com características bem distintas.

Uma das abordages poderia ser realizada por um analísta de dados, não especializado em problemas de classificação de áreas de terreno. Poderia formular o problema como um problema de inferência estatística (a inferência é realizada com base na amostra - conjunto de parcelas de classe conhecida) em que são testados os parâmetros do modelo que é utilizado para classificar a totalidade das parcelas. O modelo obtido desse modo é simples e dependente, unicamente, do modelo estatístico utilizado e dos dados que caracterizam os objectos.

Outra abordagem, basicamente diferente, seria realizada por um especialista (um foto-interpretador, um interpretador de dados de satélite, ou outro). Consideraria o domínio do problema (a área considerada, as características observadas, o conjunto de classes) de uma forma global, interrelacionando a informação disponível e tomando decisões de acordo com os dados e com o seu conhecimento. Em situações onde pareceria haver inconsistência de informação (p.e., a existência de uma zona urbana numa parcela muito declivosa) tentaria encontrar explicações (p.e., a região pode ser muito povoada, as condições ambientais podem ser mais favoráveis do que nas regiões vizinhas,...) e assim completar progressivamente o seu conhecimento sobre o problema. Esse processo de aprendizagem auxiliar-o-ia para subsequentes tomadas de decisão.

Comparando as duas perspectivas verifica-se que se situariam em pontos extremos de um referencial no qual seriam considerados dois aspectos: o automatismo do processo (máximo na abordagem estatística e mínima na classificação pericial) e a descrição da estrutura dos dados subjacente à regra de classificação (situação inversa). O desafio levantado após esta constatação é o seguinte: "Será possível desenvolver um método de classificação que permita aplicar mecanismos cognitivos semelhantes aos do perito de uma forma automatizada? "

No caso da resposta ser afirmativa resultaria uma consequência importante: os mecanismos cognitivos e o conhecimento do perito ou resultante do processo de aprendizagem seriam captados em estruturas e portanto manipuláveis (ao contrário do que acontece quando o perito, autonomamente, analisa o problema).

Existem diversas publicações onde são expostos métodos de classificação que permitem integrar informação dos dados e conhecimento pericial. Os resultados obtidos para situações reais são, em muitos casos, positivos, isto é, conduzem a níveis de qualidade de classificação superiores aos dos métodos estatísticos usualmente utilizados isoladamente como pode ser verificado, por exemplo, em Middelkoop *et al.* (1991) (incorporação de conhecimento sobre relações temporais num método convencional de classificação supervisada de dados de satélite), Skidmore (1989) (incorporação de conhecimento pericial - sobre probabilidades *a priori* das classes para as várias condições possíveis e sobre o contexto espacial - através de um método bayesiano, para classificação de manchas florestais) e Srinivasan *et al.* (1990) (utilização de um sistema pericial para classificar áreas urbanas a partir de dados de várias origens). Métodos oriundos da área da Inteligência Artificial tem sido igualmente utilizados com sucesso em problemas de classificação para a resolução dos quais não está disponível uma amostra de objectos de classificação conhecida. Alguns exemplos dessas aplicações podem ser encontrados em Mehldau *et al.* (1990) (incorporação de regras num processador convencional de imagens de detecção remota), Moller-Jensen (1990) (sistema pericial para construção de mapas de zonas urbanas a partir de dados de satélite) e Qian *et al.* (1990) (utilização do conhecimento sobre o contexto espacial na caracterização de bacias hidrográficas).

Os métodos referidos dão apenas uma resposta parcial à questão atrás formulada pois não apresentam estruturas sufientemente flexíveis de representação do conhecimento. Neste texto será proposto um método que se julga ir mais além na resposta à dita questão. Esse método é de aplicação geral a um grupo variado de problemas de classificação e não apenas ao problema descrito no início da introdução.

1.2. Enquadramento geral do tema

Os problemas de classificação (problemas de afectação de objectos a classes prédefinidas e disjuntas) apresentam uma grande generalidade, seja na gama de disciplinas às quais são aplicados (como exemplos poderiam ser referidos problemas de áreas tão diversas como a medicina, antropologia, ecologia, taxonomia, psicologia, linguística, ciências agrárias e outras), seja nas formas com que são apresentados (problemas de inferência estatística, de aprendizagem, de modelação ou de reconhecimento de padrões).

A abordagem convencional para a resolução dos problemas de classificação consiste, basicamente, em identificar os objectos a classificar como vectores de um espaço $\mathbb{R}^p$ e as classes como regiões desse espaço e em resolver um ou mais problemas de optimização por forma a determinar uma regra de classificação que minimize o erro cometido na afectação dos objectos às classes.

Essa abordagem permite, no caso das variáveis consideradas serem adequadas, obter regras de classificação eficientes, isto é, que afectam correctamente uma grande proporção dos objectos às respectivas classes (objectivo 1).

A abordagem do perito, pelo contrário, atende a outros dois objectivos que podem ser considerados importantes: tomar em consideração informação não apenas descritiva dos objectos, como seja o conhecimento de peritos sobre o domínio do problema (objectivo 2); e discernir e compreender a estrutura preditiva das variáveis que caracterizam os objectos a classificar, ou seja, modelizar o domínio dos objectos na perspectiva da classificação (objectivo 3).

O perito tem a capacidade de realizar uma análise mais profunda porque utiliza um formalismo mais complexo que o formalismo no qual se apoiam os métodos convencionais.

Considerem-se alguns aspectos que caracterizam e condicionam esse formalismo:

- conhecimento - pode considerar-se que o conhecimento envolvido num problema de classificação se situa a quatro níveis. O primeiro está ligado à representação dos objectos a classificar. Essa representação pode ser numérica (como vectores de um espaço real multidimensional) ou simbólica (forma alternativa de representação baseada numa semântica do domínio dos objectos mais adequada do que a semântica dos números). No nível da representação as entidades consideradas podem ser designadas por **conceitos**. Os segundo e terceiro níveis do conhecimento envolvem o conhecimento pericial. O segundo nível é constituído

por associações (empíricas) do tipo causa-efeito entre conceitos, as quais podem ser designadas por **regras** (p.e., um perito em ciências agrárias sabe que num regossolo a produtividade é baixa). O terceiro é constituído por entidades mais complexas, os **modelos**, que consistem em colecções de regras interrelacionadas. Finalmente, pode-se considerar um nível mais elevado de conhecimento que intervém na procura de soluções para o problema de classificação - a **estratégia**. Por estratégia entende-se o conjunto de regras e procedimentos para a exploração de todo o restante conhecimento.

- simbolismo - a representação e o processamento simbólico dos objectos formam um quadro no qual o processo de resolução do problema de classificação se pode inscrever por forma a que o conhecimento existente seja explorado eficientemente. Embora variáveis não quantitativas possam ser transformadas por forma a poder ser utilizada uma codificação numérica, a representação simbólica apresenta algumas características particulares que não são contempladas pela representação numérica como sejam: a) a informação está explicitamente contida na descrição qualquer que seja o tipo de característica (p.e., afirmar que "o objecto A tem côr azul" é mais informativo do que afirmar que "a variável CÔR toma um determinado valor numérico") o que permite facilmente estabelecer relações entre domínios de variáveis (p.e., para uma parcela de terreno caracterizada por não estar coberta por vegetação não é interessante conhecer o valor da variável que indica a espécie arbórea mais frequente); b) um "objecto simbólico" é uma descrição em compreensão de um conjunto de objectos (p.e., "ter côr azul" identifica o conjunto de objectos que têm côr azul) que pode ser ampliado - generalizando o objecto simbólico (p.e., "não ser amarelo") - ou reduzido - especializando o objecto simbólico (p.e., "ter côr azul e forma cúbica"). Pode-se tirar partido da característica *a*) e da semântica do domínio do problema para simplificar a descrição dos objectos (p.e., a especialização do objecto simbólico "solo pouco espesso" em "solo pouco espesso e de textura grosseira" pode ser expressa como "regossolo", que é um termo familiar para os especialistas em ciências agrárias).

  O processamento simbólico de dados é caracterizado por seguir os seguintes princípios (cf. Diday, 1989): a) deve ser evitada a utilização de codificações redutoras da realidade; b) o conhecimento (gerado automaticamente ou fornecido pelo perito) deve dirigir os métodos de análise dos dados; c) as respostas ao problema devem ser dadas em termos mais gerais do que os dados de partida; e d) os resultados devem ser de fácil interpretação e utilizáveis para definir

uma base de conhecimento sobre o domínio dos objectos. Este último princípio é muito importante pois garante que os resultados se expressem em termos manipuláveis pela inteligência humana o que lhes confere a característica de poderem ser compreendidos e criticados por um público vasto de especialistas (especialistas sobre o domínio dos objectos e não sobre os métodos de análise dos dados utilizados).

O processamento simbólico permite construir, com base no conhecimento existente, expressões ou estruturas simbólicas que dão resposta ao problema de classificação. Para além da criação o processo envolve, ao longo da sua execução, modificações, reproduções e destruições dessas estruturas.

- raciocínio - raciocínio pode ser definido como uma coordenação lógica de dois ou mais juízos com o fim de extrair ou demonstrar uma conclusão (Coelho *et al.*, 1992). Um raciocínio pode ser decomposto num conjunto de pequenas inferências individuais ordenadas. A cadeia ordenada de inferências realizadas ao ser desenvolvido um raciocínio para dar uma resposta ajustada a um determinado tipo de problema não deve ser definida à partida mas, ao contrário, deve depender dos dados e do conhecimento existente sobre o domínio do problema. Isto significa que um método de resolução do problema de classificação - que se pode apoiar numa vasta gama de tipos de raciocínio conhecidos e modelados - deve, preferencialmente, ter a capacidade de escolher, passo a passo, o tipo de raciocínio a utilizar e os critérios associados à tomada de decisão. Esta abordagem destaca-se da dos métodos de resolução que consistem numa sequência previamente definida de instruções que define o processo de resolução do problema em função, apenas, da instância do problema.

O conjunto de aspectos acima referidos, pela sua natureza e especificidade, permite enquadrar o problema de classificação (quando esse problema é formulado atendendo aos objectivos 1, 2 e 3, apresentados atrás) na área científica da Inteligência Artificial.

Assim sendo, podem ser utilizados métodos, técnicas e instrumentos de Inteligência Artificial e Engenharia do Conhecimento para construir um método de resolução do problema de classificação.

No que respeita à representação dos níveis de conhecimento, serão utilizados objectos simbólicos, regras de produção [1], redes semânticas e árvores de decisão. Os

[1] Regra de produção é uma instrução da forma SE ENTÃO, contendo conhecimento sobre um certo domínio de aplicação. A regra de produção pode ter associada uma medida de incerteza.

conceitos serão representados por acontecimentos elementares e asserções (objectos simbólicos simples). O segundo nível será constituido por regras de produção em que as premissas e as conclusões podem ser características dos objectos (estados das variáveis) ou classes. Os modelos utilizados serão redes semânticas. Estas estruturas são grafos nos quais os vértices simbolizam conceitos e as arestas representam a existência de uma relação directa entre dois vértices. A estratégia será suportada por uma árvore de decisão e consiste numa procura, primeiro em profundidade, das afectações a estabelecer entre objectos simbólicos e classes.

Os raciocínios podem ser processados automaticamente através de três tipos de mecanismos de inferência: a indução (geração de leis gerais a partir de exemplos), a dedução (geração de conclusões a partir de premissas) e a abdução (geração de explicações mais plausíveis). Estes três tipos de raciocínio situam-se em categorias distintas no que respeita ao grau de precisão: a dedução é um tipo de raciocínio rigoroso, baseado na lógica matemática; a indução e a abdução são raciocínios não rigorosos, também designados por raciocínios naturais. Dessa diferença resulta uma consequência metodológica importante: ao invés do raciocínio baseado na lógica clássica em que as crenças não podem ser revistas, nos sistemas de raciocínio que utilizam modos de raciocínio naturais as inferências realizadas numa dada fase de resolução do problema podem ser rejeitadas e eventualmente substituidas por outras em fases posteriores.

Estabelecida a forma de representação e organização do conhecimento e os modos de o explorar (mecanismos de inferência ou raciocínios) o método de resolução do problema é projectado por forma a serem respeitados os princípios do processamento simbólico atrás enunciados.

Pelo facto do modelo conceptual do método ser oriundo da área da Inteligência Artificial, isso não significa que não devam ser considerados conceitos, métodos e técnicas de outras áreas sempre que essas soluções forem vantajosas em função dos objectivos fixados à partida. No caso concreto da construção do método de classificação proposto neste trabalho, a melhor solução encontrada para manipular estruturas de representação do conhecimento consideradas baseia-se em técnicas de Probabilidades (para a incerteza sobre o conhecimento) e de Investigação Operacional (para estruturas em grafos).

A tentativa de integração de diversos métodos matemáticos por forma a construir um bom método para a resolução de um problema relativamente complexo, como é o caso do problema considerado, foi, aliás, um factor de motivação para a realização deste trabalho.

1.3. Descrição do trabalho

No capítulo 2 serão descritos os conceitos básicos, a informação disponível e os objectivos do problema. É também apresentada uma breve revisão dos principais métodos de classificação. O capítulo 2 inclui ainda uma breve descrição do processo de classificação (que é, essencialmente, interactivo) e um pequeno exemplo para ilustrar as noções abordadas.

No capítulo seguinte é descrita, conceptualmente, a estratégia de resolução do problema, deixando-se para os capítulos 4 e 5 a descrição das principais componentes operacionais do método. No capítulo 4 serão apresentadas técnicas baseadas numa árvore de representação hierárquica dos objectos da amostra e em 5 será descrito um modelo gráfico e probabilístico de representação do conhecimento e um conjunto de técnicas que permitem manipular e explorar esse modelo. Em 6 é descrita a integração de todas as componentes anteriormente apresentadas segundo o modelo conceptual proposto em 3. No capítulo 6 será também discutida a satisfação dos objectivos propostos e será analisada a implementação do método segundo vários aspectos (arquitectura, complexidade computacional e linguagem de implementação).

Um exemplo real será apresentado e resolvido no capítulo 7. Por forma a tornar mais interessante a análise, o método proposto é comparado com dois métodos que já provaram a sua eficiência em classificação. A comparação será realizada relativamente aos três objectivos apontados. Em 7 é possível encontrar também uma sequência de passos do método de classificação proposto e a análise do processo de interacção entre um perito e o programa.

Finalmente, no capítulo 8 serão apresentados alguns comentários sobre o conjunto deste trabalho.

2 . Definição formal do problema e dos objectivos. Principais métodos de classificação. Caracterização geral do método proposto

2.1. Definições prévias

Gnanadesikan *et al.* (1989) distinguem os problemas de classificação em dois grandes grupos. O primeiro inclui problemas de classificação para os quais as classes estão definidas à partida e para os quais se dispõe de uma amostra de objectos de classificação conhecida (amostra de treino). O objectivo é classificar correctamente os objectos cuja classe é desconhecida. Este grupo é identificado por **classificação supervisada** ou *discriminant analysis*. No segundo grupo de problemas o objectivo é determinar classes para os objectos a partir das suas características. Este grupo é identificado por **classificação não supervisada** ou *cluster analysis*.

Como foi referido na introdução o grupo que irá ser considerado é o grupo de problemas de classificação supervisada.

Seja N o número de objectos a classificar e seja p o número de características (quantitativas ou qualitativas, e neste caso ordinais ou nominais) que descrevem cada um dos objectos.

Denote-se por $\Omega = \{O_1, ..., O_N\}$ o conjunto dos objectos. Seja D_k o conjunto de valores que a k-ésima característica pode tomar. A cada conjunto D_k associe-se uma partição finita $P_k = \{P_{k_1}, ..., P_{k_{n_k}}\}$ com n_k elementos. Designe-se variável (V_k) a uma aplicação de Ω em D_k e seja $V_k(O_j)$ o valor (que pode ser numérico ou simbólico) da k-ésima característica do objecto O_j. São consideradas, portanto, p variáveis. As **modalidades** da variável V_k são os elementos de P_k. Os **atributos** de um objecto são as modalidades que o caracterizam isto é, o objecto O_j tem o atributo P_{k_h} se $V_k(O_j) \in P_{k_h}$. Um objecto O_j pode ser univocamente representado por um p-uplo $\mathbf{x} = (v_1, ..., v_p)$, sendo $v_k = V_k(O_j)$ para todo o k. O conjunto de todos os p-uplos possíveis designa-se **domínio dos objectos** (Ω').

As definições anteriores permitem, por sua vez, definir **objectos simbólicos**. Designa-se **acontecimento elementar** um objecto simbólico definido pela expressão $[V_k \in P_{k_h}]$. A **extensão** do acontecimento elementar $[V_k \in P_{k_h}]$ é o conjunto de objectos $\{O_j : V_k(O_j) \in P_{k_h}\}$. Uma **asserção** é uma conjunção de acontecimentos elementares e representa-se por $a = [v'_1 \in P'_{1_{h_1}}] \wedge ... \wedge [v'_q \in P'_{q_{h_q}}]$, sendo $q \leq p$, e a sua extensão em Ω, $|a|_\Omega$, é o conjunto de objectos de Ω que verificam a conjunção de

expressões que definem a asserção, isto é, $|a|_\Omega = \{\omega \in \Omega : v_i'(\omega) \in P'_{i_{h_i}}, \forall_{i'=1,...,q}\}$. Os acontecimentos elementares e as asserções são, segundo Diday (1989), os objectos simbólicos mais simples.

São prédefinidas m classes ($C = \{c_1, ...c_m\}$) disjuntas. Vamos supor que cada um dos N objectos a classificar pertence a uma e uma só classe (seja C a função injectiva do conjunto dos objectos no conjunto das classes, sendo, portanto, $C(O_j)$ a classe a que pertence o objecto O_j). Sendo assim o conjunto das m classes forma uma partição do domínio dos objectos.

O problema de afectação dos objectos resolve-se através de uma regra de decisão. Uma regra de decisão é uma função que define uma partição $P = \{R_1, ..., R_m\}$ do espaço Ω'. Diz-se que o objecto O_j representado em Ω' por $\mathbf{x}$ pertence à classe c_i se $\mathbf{x} \in R_i$ (Hand, 1981), isto é, uma regra de decisão afecta uma classe c_ia cada conjunto R_i. Pode-se generalizar facilmente a noção de regra de decisão a uma função que define uma partição com um número de elementos superior ao número de classes, desde que seja possível reduzir essa função ao caso anterior através de uniões de elementos da partição, ou seja, $P' = \{R'_1, ..., R'_M\}$, com $M \geq m$, define uma regra de decisão se $\forall_{i\in\{1,...,M\}}$, $\forall_{\mathbf{x}\in\Omega}$, $\exists_{k\in\{1,...,m\}} : [\mathbf{x} \in R'_i \Rightarrow C(\mathbf{x}) = c_k]$.

Cada um dos elementos da partição (P') será designado **região de decisão**. A expressão **regra de afectação** designará a aplicação de P' em $\{c_1, ..., c_m\}$. A expressão **regra de classificação** designará a definição das regiões de decisão e da regra de afectação. Designar-se-á por **afectação** uma relação estabelecida entre uma região de decisão e uma classe. A regra de afectação é, portanto, o conjunto das afectações.

2.2. A informação disponível

Apresentam-se, neste ponto, as diversas componentes da informação disponíveis para resolver o problema de classificação.

- A **amostra** (também designada por amostra de treino porque podem ser induzidas, com base nessa amostra, por vezes através de processos iterativos, regras de classificação) é um conjunto de L objectos cuja classe é conhecida à partida, isto é, $\{O_1, ..., O_L\}$ tais que $C(O_j)$ tem valor conhecido, $\forall_{j\in\{1,...,L\}}$. Os objectos que constituem essa amostra podem ser escolhidos de várias formas mas devem, na medida do possível, seguir os dois seguintes princípios: cobrirem

o conjunto das classes e estarem bem distribuídos pelo domínio dos objectos.

- As **características dos objectos** a classificar, ou seja, os valores que tomam as variáveis $\{V_1, ..., V_p\}$ para cada objecto que se pretende classificar.

- **Conhecimento pericial**. O conhecimento pericial inclui toda a informação disponível não referida atrás que pode contribuir para a resolução do problema. É introduzida sob forma de *hipóteses sobre relações entre as classes, os objectos e as variáveis que os descrevem*, hipóteses essas que traduzem o conhecimento do perito sobre o domínio dos objectos e as classes consideradas e que explicam relações verificadas na amostra.

2.3. Os objectivos do método proposto

Como foi referido na introdução, o método de classificação proposto visa, basicamente, três objectivos.

O primeiro consiste em definir uma regra de classificação que permita classificar, cometendo o menor número de erros possível, a totalidade dos objectos. O método de classificação proposto neste trabalho permite tomar em consideração diversos tipos de informação. Consequentemente, tem-se a pretensão de conseguir construir uma regra de classificação mais correcta (porque tem como suporte uma maior quantidade de informação) do que uma regra que seria unicamente induzida a partir da amostra.

A satisfação do primeiro objectivo pode ser quantificada (Hand, 1981) pela taxa de erro (número de objectos mal classificados/número total de objectos) que pode ser real (calculada sobre um conjunto de objectos que não pertencem à amostra de treino) ou aparente (para um conjunto de objectos que inclui objectos que pertencem à amostra de treino). Na maior parte dos casos, evidentemente, apenas pode ser calculada a taxa aparente que diz respeito à própria amostra. É necessário tomar precauções ao utilizar este último indicador pois uma taxa demasiado baixa pode significar que a regra de classificação é demasiado ajustada à amostra e que não é generalizável a outros objectos.

O segundo objectivo consiste em tirar partido de toda a informação disponível. Em particular, pretende-se que seja tirado partido do conhecimento pericial para a construção da regra de classificação.

O terceiro consiste em captar a estrutura preditiva das variáveis que caracterizam os objectos a classificar de uma forma que permita definir uma base de

conhecimento para o domínio dos objectos. Pretende-se que essa base de conhecimento seja de *fácil interpretação* e, portanto, é necessário que as estruturas de representação do conhecimento constituam um suporte *transparente* da informação.

2.4. Breve apresentação dos principais métodos de classificação

Justifica-se fazer uma breve apresentação dos principais métodos utilizados para classificação supervisada por duas ordens de razões: 1) para ilustrar a diversidade de formas de regra de classificação que podem resultar; 2) para poder comparar o método de classificação que será exposto ao longo deste trabalho com métodos que já provaram ser eficientes na resolução de problemas de classificação supervisada. Em relação e este último aspecto convém, desde já, realçar que a comparação deverá ser realizada em relação aos vários objectivos considerados para o método proposto (v. 2.3). Quando se refere que os métodos que serão apresentados em seguida provaram ser eficientes, isso significa apenas que conduzem à construção de uma regra que classifica bem os objectos de classe desconhecida (primeiro objectivo considerado), não tendo nenhum significado em relação aos outros dois objectivos.

O conjunto de métodos apresentados em seguida não é, de forma alguma, exaustivo. É retirado de uma revisão realizada pelos membros do "Panel on Discriminant Analysis, Classification and Clustering" apresentada em (Gnanadesikan *et al.*, 1989).

Os grandes grupos de métodos de classificação diferem na forma de representação da regra de classificação.

Seja $\mathbf{x}$ o p-uplo de características que representa um objecto (se as características forem caracterizadas por valores numéricos, $\mathbf{x} \in \mathbb{R}^p$) e $\{c_j, j = 1, ..., m\}$ o conjunto de classes consideradas.

Uma das principais regras de classificação é a **regra de erro mínimo de Bayes** que define as regiões de decisão da seguinte forma:

$$P(c_k|\mathbf{x}) \geq P(c_j|\mathbf{x}),\ \forall_{j \neq k} \Rightarrow C(\mathbf{x}) = c_k.$$

Como se pode verificar a regra baseia-se no conhecimento de distribuições teóricas de probabilidade [1]. Um outro tipo de regra muito utilizado baseia-se numa *função*

[1] Obtem-se uma regra equivalente mas mais conveniente se se aplicar o teorema de Bayes: $P(c_i|\mathbf{x}) = \frac{p(\mathbf{x}|c_i) \cdot P(c_i)}{p(\mathbf{x})}$.

discriminante e tem a forma:

$$q_k(\mathbf{x}) \geq q_j(\mathbf{x}),\ \forall_{j \neq k} \Rightarrow C(\mathbf{x}) = c_k,$$

sendo q_i uma determinada função do vector de características $\mathbf{x}$. Nos dois casos apresentados as superfícies de decisão são dependentes das funções envolvidas (as distribuições de probabilidade podem ser normais ou outras, as funções discriminantes lineares, quadráticas ou outras) e definem, geralmente, regiões de decisão convexas.

É habitual distinguir os métodos paramétricos dos métodos não paramétricos de classificação. Nos métodos paramétricos supõe-se, à partida, que as funções desconhecidas têm uma determinada expressão e estimam-se os parâmetros dessas funções com base na amostra. Os diversos métodos paramétricos diferem nas expressões teóricas admitidas e nos métodos de estimação dos parâmetros que podem[2] ser de máxima verosimilhança, minimização de medidas de distância, método de Bayes e métodos sequenciais para os parâmetros das distribuições de probabilidade e de programação linear, gradiente, critério de Fisher e regrassão linear múltipla para os parâmetros das funções discriminantes. Em determinadas condições teóricas da população dos objectos prova-se que alguns dos métodos paramétricos são equivalentes. Os métodos não paramétricos podem ser utilizados para estimar a função de distribuição de probabilidade e, dessa forma, construir uma regra de erro minimo de Bayes (estão neste caso[3] os estimadores de *kernel*, o método dos k vizinhos mais próximos, os métodos baseados em séries e no modelo de regressão múltipla). Em todos os métodos atrás referidos os objectos devem ser representados por um vector numérico, o que implica converter as variáveis não numéricas em variáveis numéricas (binárias, geralmente).

Um grupo de métodos de classificação não paramétricos que é habitualmente considerado separadamente suporta-se em *árvores*. A árvore fornece uma representação do tipo hierárquico do espaço dos objectos. É também habitual distinguir as árvores binárias (que são descritas em pormenor em Breiman *et al.*, 1984) das árvores não binárias. Essas estruturas diferem no número de ramificações em cada nó da árvore. Entre as árvores não binárias destaca-se o método ID3 (Quinlan, 1986) que consiste na construção de uma árvore de classificação induzida pela amostra de treino. Os nós dessa árvore representam asserções. O método é, portanto, adequado para objectos caracterizados por um número finito de atributos, o que implica a

[2]Segundo Hand (1981).
[3]Segundo Hand (1981) e Tomassone *et al.* (1988).

discretização das variáveis numéricas contínuas envolvidas.

No que respeita ao terceiro objectivo referido em 2.3, os utilizadores de métodos baseados em árvores defendem que a regra de classificação obtida através desses métodos é mais facilmente explicativa, em relação à estrutura do domínio do problema, do que as regras de erro mínimo de Bayes ou baseadas numa função discriminante. No entanto, quando as árvores de classificação induzidas pela amostra são de grande dimensão esse argumento deixa de fazer muito sentido.

Na figura 2.1. apresenta-se um esquema sintético dos principais métodos de classificação supervisada.

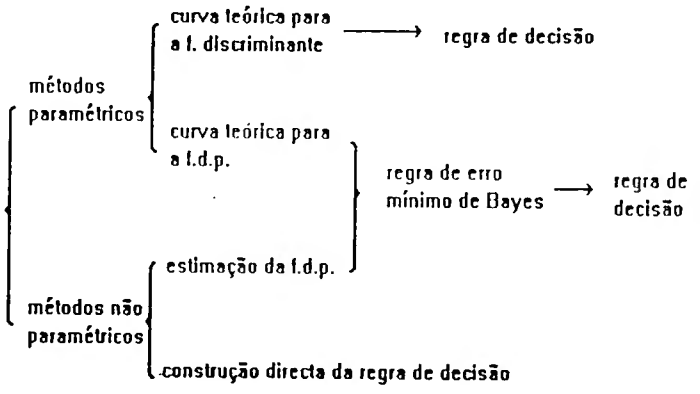


Figura 2.1 - Principais métodos de classificação supervisada

A título de exemplo refira-se que Stockwell *et al.* (1990) compararam o desempenho de vários métodos de classificação num problema real. As taxas de erro reais obtidas ordenaram, por ordem crescente de erro, os métodos: árvore de decisão induzida, CART (árvore binária), sistema pericial (baseado em regras), modelo linear (estimado por regressão linear) e análise de componentes principais. O que se pode concluir por estes resultados comparativos e pela eficiência conseguida em variadíssimas aplicações realizadas com métodos estatísticos, é que, à partida, não existem "bons" métodos e "maus" métodos. O desempenho de cada um depende das características do problema e dos objectivos que se pretende atingir.

Relativamente ao método proposto neste trabalho, a forma de regra de classificação adoptada, por um conjunto de razões que serão apresentadas no capítulo seguinte, é a de árvore de classificação. Mais concretamente, a regra de classificação obtida através do método proposto é constituída por um conjunto de regras de produção, nas quais as premissas são asserções e as conclusões são as classes. Esse conjunto de regras é representado através de uma árvore de decisão na qual cada nó

representa uma asserção.

A árvore tem duas funções: o suporte da regra de classificação e o suporte da estratégia de resolução do problema de classificação. Esta dupla função da árvore de decisão é uma característica essencial do método proposto.

2.5. O processo de classificação

O processo de classificação designará o processo de resolução do problema de classificação, de modo a atingir os três objectivos referidos em 2.3.

A processo de classificação é composto, por uma sequência de passos que canalizam a informação disponível desde um estado *bruto* até à um conjunto de estruturas que dão resposta ao problema de classificação. O conhecimento captado ao longo desse processo é registado. A regra de classificação obtida é , finalmente, utilizada para classificar a totalidade dos objectos com base nas suas características.

As características particulares do processo de classificação associado ao método de classificação proposto neste trabalho são:

- a regra de classificação é obtida através de um processo interactivo no decorrer do qual é integrado o conhecimento do perito e a informação contida na amostra de treino;
- o conhecimento resultante desse processo interactivo é armazenado em diversas estruturas e explorado nas fases subsequentes do processo;
- as decisões podem ser tomadas em situações de incerteza e de inconsistência da informação;
- o problema pode ser dividido em subproblemas restritos a situações mais homogéneas que o problema inicial;
- o processo de classificação fornece ao perito uma base de referência que, para além de lhe permitir formalizar o seu conhecimento e controlá-lo, abre-lhe novas vias de raciocínio;
- um conjunto das estruturas de representação do conhecimento caracterizam, no final do processo de classificação, o domínio dos objectos.

2.6. Exemplo

Apresenta-se, em seguida, um pequeno exemplo para ilustrar algumas noções definidas neste capítulo como sejam os acontecimentos elementares e as asserções como representações dos objectos, a extensão de uma asserção, a regra de classificação, a qualidade dessa regra e o conhecimento pericial envolvido no problema exemplificativo.

O exemplo é retirado de um problema real que será analisado em pormenor no capítulo 7. É um problema de afectação de parcelas de terreno a uma de duas classes. Uma das classes corresponde à presença de uma determinada espécie animal, a outra corresponde à ausência. O interesse prático desta aplicação é o da previsão do impacte ecológico, sobre a espécie, da construção de um observatório astronómico, ou seja, pretende-se prever, nas parcelas de terreno onde se pretende construir o observatório, se existe ou não a espécie considerada e quantificar, portanto, o efeito directo sobre a espécie da construção do edifício.

Considere-se que os objectos considerados são parcelas de terreno e são caracterizados por: a) altitude; b) densidade de copado - proporção da área da parcela coberta por copas de árvores; e c) nível alimentar - índice da qualidade e quantidade de alimento disponível na parcela (principalmente frutos) para a alimentação de um determinado animal.

A questão que se põe é a da presença ou ausência de nichos do referido animal na parcela. O problema de afectação coloca-se, então, da seguinte forma: *sabendo os valores das variáveis altitude, copado e alimento para uma determinada parcela de terreno, é possível prever a presença ou ausência de nichos do animal nessa parcela?*

Como é evidente, neste problema existem duas classes que estão definidas à partida. Pode associar-se aos domínios das variáveis partições finitas e assim definir modalidades para essas variáveis. Por exemplo, pode considerar-se que *altitude* tem as modalidades [*altitude* $< 10^4$ ft] e [*altitude* $\geq 10^4$ ft], que *alimento* tem as modalidades [*alimento* = *elevado*], [*alimento* = *médio*] e [*alimento* = *baixo*] e que *copado* tem as modalidades [*copado* = *denso*] e [*copado* = *esparso*]. Utilizando a notação definida em 2.1, pode-se formalizar a informação acima do seguinte modo:

C={presença,ausência}, e

variáveis (V_k):	altitude	alimento	copado
domínio (D_k):	[0 ft,20000 ft]	{nulo,baixo,médio,elevado}	[0%, 100%]
partição associada (P_k):	$\{< 10^4$ ft, $\geq 10^4$ ft$\}$	{baixo,médio,elevado}	{esparso,denso}

Considere-se o seguinte conjunto de 10 objectos cuja classificação é conhecida e que constituem, por isso, uma amostra.

atributos e classes dos objectos da amostra				
objectos	altitude	alimento	copado	classe
parcela 1	$< 10^4$ ft.	médio	esparso	ausência
parcela 2	$\geq 10^4$ ft.	elevado	denso	presença
parcela 3	$< 10^4$ ft.	médio	denso	ausência
parcela 4	$< 10^4$ ft.	elevado	denso	presença
parcela 5	$< 10^4$ ft.	médio	esparso	ausência
parcela 6	$\geq 10^4$ ft.	elevado	denso	presença
parcela 7	$< 10^4$ ft.	médio	denso	ausência
parcela 8	$\geq 10^4$ ft.	elevado	denso	ausência
parcela 9	$\geq 10^4$ ft.	elevado	esparso	presença
parcela 10	$< 10^4$ ft.	médio	esparso	ausência

Tabela 2.1 - Amostra ilustrativa de 10 objectos

Um exemplo de objecto simbólico é $[alimento = médio]$. A sua extensão é constituída por todos os objectos que têm esse atributo, ou seja, se $a = [alimento = médio]$ então $|a|_{amostra} = \{O_1, O_3, O_5, O_7, O_{10}\}$. O objecto simbólico $[copado = denso] \wedge [alimento = médio]$ é uma asserção e a sua extensão é constituída por todos os objectos que têm, simultaneamente, esses dois atributos, ou seja, $\{O_3, O_7\}$.

Para este conjunto de objectos pode ser construída uma regra de classificação. Sendo os objectos descritos por variáveis categoriais, as regiões de decisão são, naturalmente, definidas por asserções. Por exemplo a regra de classificação poderia ser:

- $O_j \in |[altitude < 10^4 ft.]|_{amostra} \Rightarrow C(O_j) = ausência.$
- $O_j \in |[altitude \geq 10^4 ft.]|_{amostra} \Rightarrow C(O_j) = presença$

Esta regra permitiria classificar correctamente 8 objectos da amostra e, portanto, teria uma taxa de erro aparente sobre a amostra de 20%.

O conhecimento diponível para classificar os objectos é constituído pela tabela 2.1, pelos atributos de todos os objectos que se pretende classificar e pelo conhecimento pericial existente sobre o domínio dos objectos. Este último poderia ser representado na forma de hipóteses tais como:

- *as condições de desenvolvimento das copas são mais favoráveis em zonas de altitude elevada*;
- *a densidade do copado favorece a produção de alimento*; ou
- *boas condições de copado e de alimento favorecem a presença do animal.*

A estruturação do conhecimento pericial e a sua utilização ao longo do processo de classificação serão descritas no decorrer deste trabalho.

3 . Estratégia do método de resolução do problema

Neste capítulo a resolução do problema será definida como uma procura de uma solução num espaço de soluções admissíveis. A solução procurada é uma solução, pertencente ao conjunto de soluções admissíveis, que dá uma boa resposta ao problema, isto é, uma solução que satisfaz um determinado critério, definido de acordo com o(s) objectivo(s) da resolução do problema.

Para definir o método de classificação é, então, necessário definir: 1) as soluções admissíveis; 2) o critério que distingue as "boas" soluções; 3) os operadores que permitem determinar uma solução admissível a partir de outra solução admissível; e, para evitar que o processo de procura se torne muito longo, 4) regras para orientar, de forma eficiente, a procura.

Esses aspectos irão ser discutidos ao longo deste capítulo, assim como a representação do espaço de procura e a representação da informação incerta envolvida no problema. Relativamente às regras de orientação da procura e ao critério que define as soluções procuradas, neste capítulo será apenas feito um enquadramento geral. O desenvolvimento destas questões será apresentado nos capítulos seguintes.

3.1. Definição do método de resolução do problema como uma procura num espaço de estados

3.1.1. O espaço das soluções do problema

Embora tenha sido considerado que o método de classificação proposto tem diversos objectivos, uma solução admissível do problema designará simplesmente uma regra de classificação, ou seja, uma aplicação de um determinado conjunto de regiões de decisão (P') no conjunto das classes definidas à partida (C). Uma solução admissível poderá, indistintamente, ser designada por estado do problema.

Considere-se, a partir deste momento, que as regiões de decisão são definidas por asserções e, portanto, que P' é um conjunto de asserções que constitui uma partição de Ω'.

Dessa suposição resulta que o espaço de estados onde uma solução do problema pode ser encontrada tem um número finito de elementos (pois, por hipótese, as partições dos domínios das variáveis são finitas e o número de classes também é

finito) e é constituído por todas as possíveis aplicações de $\mathcal{P}'$ em $\mathcal{C}$, para todos os $\mathcal{P}'$. O número de soluções admissíveis cresce com o número de variáveis, com o número de modalidades das variáveis e com o número de classes.

3.1.2. A redução do problema a subproblemas

As soluções de um problema podem ser procuradas no espaço de soluções admissíveis pela redução do problema a subproblemas e pela resolução individual de cada subproblema. Cada subproblema criado deve ser de resolução mais simples que o problema inicial.

O processo de decomposição para o problema de classificação consiste, para um determinado estado, em considerar isoladamente cada uma das regiões de decisão definidas nesse estado e resolver o problema de classificação associado a cada uma dessas regiões. Em particular, considerando como estados possíveis do problema os estados definidos em 3.1.1, uma solução do problema de classificação em Ω' é dada pelas soluções dos problemas de classificação colocados em cada uma das asserções que formam uma partição de Ω'.

3.1.3. A representação do espaço de procura: uma árvore de classificação

A representação utilizada põe em evidência o conjunto de subproblemas considerado em cada passo do processo de resolução do problema e define, igualmente, o estado corrente e os estados anteriormente explorados.

Essa representação consiste numa árvore (que será designada **árvore de classificação** para realçar o facto de suportar a regra de classificação).

Formalmente, uma **árvore** é um grafo sem ciclos. Um **grafo** é constituído por um conjunto de nós e por um conjunto de arestas. Cada aresta de um grafo é um par de nós. Um **caminho** entre dois nós, α e β, é uma sucessão de nós $\alpha = \gamma_0, \gamma_1, \gamma_2, ..., \gamma_n = \beta$ tal que (γ_i, γ_{i-1}) pertence ao conjunto de arestas do grafo, $\forall_{i \in \{1,...,n\}}$. Um **ciclo** é um caminho que tem como extremos o mesmo nó.

Os nós da árvore de classificação são asserções (representam conceitos). O primeiro nó criado (nó raiz) é a asserção vazia ([ ]) e representa o domínio do problema (Ω'). A ramificação de um nó da árvore processa-se da seguinte forma. Seja a uma asserção (um nó da árvore) definida pelo conjunto A de variáveis e seja V_e uma variável não pertencente a A. Então, a é ramificado por V_e no conjunto de nós $a \wedge [V_e \in P_{e_1}], ..., a \wedge [V_e \in P_{e_{n_e}}]$. Cada nó criado pode ser ramificado de forma

idêntica por uma nova variável.

Um nó não ramificado designa-se nó terminal. Represente-se por T o conjunto de nós terminais de uma determinada árvore. T é, portanto, uma partição de Ω'.

Uma aplicação de T em $\mathcal{C}$ é, de acordo com 3.1.1, um estado do problema. Caso a aplicação não esteja definida ou esteja apenas parcialmente definida, T representa um conjunto de estados do problema, o conjunto das aplicações possíveis de T em $\mathcal{C}$.

A árvore representa, por um lado, o conjunto de subproblemas considerados (o conjunto de elementos de T) e, por outro lado, o estado (ou conjunto de estados) correntemente explorado(s) do espaço de procura (o próprio T e a aplicação, de T em $\mathcal{C}$, que pode não estar totalmente definida).

3.1.4. Os operadores de mudança de estado

Os operadores de mudança de estado são operadores que transformam o estado correntemente explorado no seguinte. De modo a poder utilizar a representação em árvore descrita em 3.1.3 serão, para o método proposto, considerados os dois seguintes tipos de operadores:

- ramificação de um dos elementos de T, originando um novo conjunto de nodos terminais da árvore;
- alteração ou definição (total ou parcial) da aplicação de T em $\mathcal{C}$.

3.2. A estratégia de procura

Como foi referido anteriormente, o espaço de estados pode ter um número de elementos exponencialmente crescente com o número de variáveis e de modalidades consideradas. Por essa razão, um qualquer método de procura no espaço das soluções é sensível a uma explosão combinatória.

Uma forma de executar o controle da procura consiste na utilização de heurísticas. Uma **heurística** é uma técnica para aumentar a eficiência da procura que não garante que seja encontrada a melhor solução para o problema.

Podem considerar-se 2 tipos de heurísticas (cf. Rich, 1983):

- Heurísticas de aplicação geral. São estratégias de controle da procura aplicáveis a uma grande diversidade de problemas. Estas heurísticas são também desig-

nadas por *métodos fracos* de resolução de problemas porque não exploram o conhecimento específico do domínio do problema.

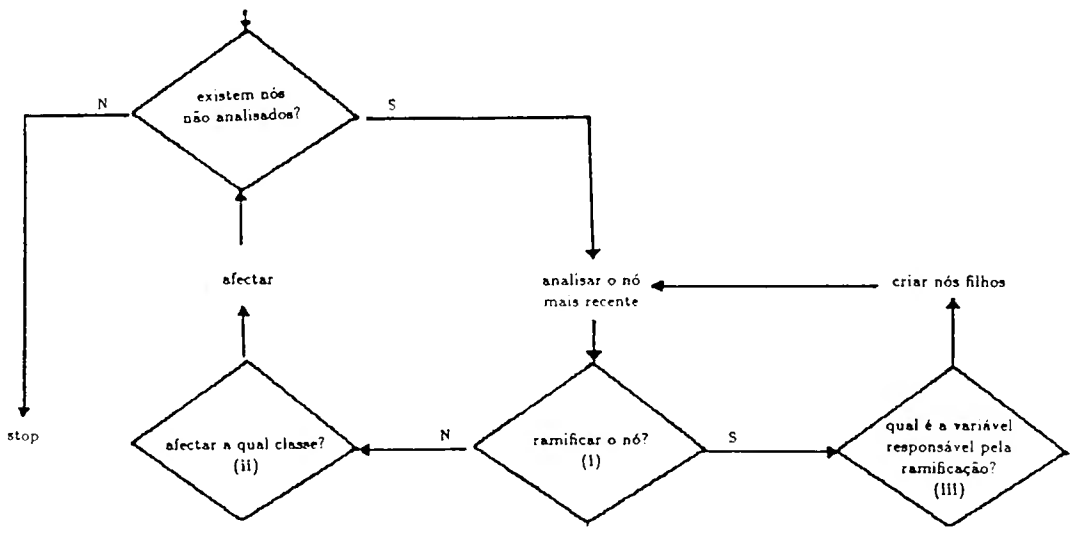
- Heurísticas específicas. São técnicas baseadas em conhecimento específico e relevante para um determinado problema. Estas últimas podem ser incorporadas no processo de procura de duas formas: 1) nas regras que definem a direcção de procura; 2) numa função heurística que avalia a "qualidade" de cada solução do problema.

3.2.1. Heurísticas de aplicação geral

No método proposto são utilizadas duas estratégias de aplicação geral: a, já anteriormente referida, redução de problemas a subproblemas e a procura primeiro em profundidade (um caso particular da estratégia de gerar e testar soluções). Frequentemente, essas duas técnicas são associadas (Kowalski, 1979).

Como foi descrito em 3.1.2, a redução do problema a subproblemas permite reformular o problema inicial como um conjunto de subproblemas mais simples. A solução do problema inicial é formada pelas soluções do conjunto de subproblemas criados.

A procura em profundidade é a estratégia utilizada para a escolha dos subproblemas a considerar. Esta técnica consiste em explorar em primeiro lugar o nó criado mais recentemente. Em relação à árvore de classificação utilizada neste trabalho, esta técnica conduz à execução do algoritmo 3.1, apresentado em seguida, para a construção da árvore.



Algoritmo 3.1 - Procura primeiro em profundidade

Em 6.1 será descrito um tipo de situação encontrada no processo de procura da solução para o qual o método de classificação proposto utiliza uma estratégia diferente da procura primeiro em profundidade.

3.2.2. Heurísticas específicas

Como foi repetidamente referido, pretende-se que método de classificação proposto neste trabalho permita explorar o conhecimento disponível sobre o domínio do problema.

Nesse sentido são construídas heurísticas que, por um lado, dirigem a procura de soluções e, por outro lado, permitem avaliar a "qualidade" de qualquer solução do problema de classificação.

É conveniente distinguir, no grupo de heurísticas específicas do domínio do problema, dois subgrupos. Por um lado, as que são construídas com base na informação contida na amostra. Por outro, as que resultam do conhecimento pericial existente. A construção destas últimas heurísticas implica a prévia integração do conhecimento pericial em estruturas que o tornem manipulável. A estrutura utilizada para o efeito é uma rede semântica. O formalismo para representação da incerteza associada a esse modelo é o da teoria das probabilidades e, por isso, a rede semântica toma a forma particular de um grafo bayesiano.

Devido à importância que as heurísticas específicas têm neste trabalho, elas terão uma descrição mais aprofundada. No capítulo seguinte será considerada a contribuição da informação da amostra e, no capítulo 5, a contribuição do conhecimento pericial.

No capítulo 6 será descrita a forma de integrar todo esse conjunto de heurísticas.

3.3. Formalização da incerteza

O conjunto de informação disponível para resolver o problema envolve informação incerta. Não existe certeza sobre a qualidade de uma determinada solução do problema de classificação (pois apenas se dispõe de uma amostra de objectos). Não existe, igualmente, certeza no que diz respeito ao conhecimento pericial.

Existem formas alternativas de descrição da incerteza, numéricas e não numéricas. Entre as formas numéricas destacam-se, habitualmente, a teoria das probabilidades, a lógica vaga e as funções de crença. A formalização da incerteza por teorias diferentes da teoria clássica das probabilidades tem vindo a ser utilizada em diversas

aplicações da área da Inteligência Artificial.

Para as necessidades de representação da incerteza que se colocam para a construção do método de classificação, a teoria das probabilidades fornece um formalismo que se considera adequado.

O interesse pelo julgamento probabilístico em Inteligência Artificial tem sido motivado, pelo menos parcialmente, pelo interesse de incorporar probabilidade em regras de produção.

Um conjunto de regras de produção (designado também por sistema de produção) é um modelo atractivo para representar a inteligência. Muitos esforços foram realizados para utilizar probabilidades em sistemas de produção (Shafer, 1987). Diversos resultados que resultam desses esforços irão ser utilizados, no presente trabalho, para a construção do método de classificação.

3.4. Complexidade de uma solução do problema

A noção de complexidade que será definida em seguida nada tem a ver com a noção de complexidade computacional. Complexidade de uma regra de classificação é definida como o número de regiões de decisão que a regra define, ou seja, $card(\mathcal{P}')$.

A noção de complexidade é importante porque está associada ao terceiro objectivo (v. 2.3) do método de classificação, nomeadamente, está associada à interpretabilidade da regra de classificação. Em geral, quanto menor fôr a complexidade mais simples será a interpretação, em termos do domínio do problema, da regra de classificação.

A noção de complexidade estará, portanto, subjacente à estratégia de resolução do problema.

4 . Heurísticas baseadas na amostra

Antes de definir as heurísticas utilizadas é necessário explicar como se pode relacionar a informação contida na amostra com os estados do problema e, particularmente, com as asserções que definem os subproblemas considerados.

Os nós da árvore são asserções (v. 3.1.3) e cada nó representa um conjunto de objectos (a extensão na amostra da asserção) e um conjunto de variáveis (as variáveis que não definem a asserção e que, portanto, estão ainda disponíveis para a ramificação do nó). A cada variável está associada um partição finita que define as modalidades da variável.

O nó que representa a totalidade da amostra é, portanto, o nó raiz. Um nó α diz-se *filho* de um nó β (reciprocamente β é designado *pai* de α) se o conjunto de objectos representados por α for um subconjunto dos objectos representados por β e se α tiver associadas todas as variáveis associadas a β com excepção de uma variável.

Na raíz da árvore situa-se a amostra, o conjunto de variáveis consideradas e, para cada uma delas, a definição de uma partição do seu domínio. O nó raíz ramifica-se da seguinte forma: é escolhida uma das variáveis disponíveis; é criado um nó *filho* por cada elemento da partição do domínio da variável escolhida; em cada nó *filho* é colocado o subconjunto dos elementos da amostra de treino que pertencem ao elemento da partição correspondente ao nó, o conjunto de variáveis do nó *pai* com excepção da variável escolhida e as partições correspondentes a esse conjunto de variáveis. A ramificação de cada nó *filho* processa-se de igual forma. Uma consequência deste tipo de ramificação é que qualquer par de nós da árvore entre os quais não haja relações de descendência (ou ascendência) é, seguramente, disjunto (não tem objectos comuns).

Utilizando as notações definidas em 2.1 pode apresentar-se formalmente o que foi descrito acima. Um objecto (O_j) tem, no que diz respeito à variável V_k, n_k atributos diferentes, cada um dos quais definido por uma das condições $[V_k(O_j) \in P_{k_1}], ..., [V_k(O_j) \in P_{k_{n_k}}]$. Ao nó raiz correspondem os L objectos da amostra e as p variáveis. É escolhida uma das variáveis associadas ao nodo, V_e, para construir a ramificação que será constituída por n_e nós *filhos*. Ao primeiro nó *filho* corresponde o conjunto de objectos $\{O_j : V_e(O_j) \in P_{e_1}\}$ [1] e o conjunto de variáveis $\{V_1, ..., V_p\} - \{V_e\}$, ao segundo o conjunto de objectos $\{O_j : V_e(O_j) \in P_{e_2}\}$ e o mesmo conjunto

[1] Qualquer conjunto de objectos associados a um nó da árvore, com excepção da *raiz*, será designado por **subamostra**.

de variáveis, e assim sucessivamente até esgotar os elementos de P_k. Este processo é repetido para cada nó que é ramificado.

4.1. Heurística para escolha da ramificação da árvore de classificação

Uma heurística, baseada na teoria da informação, será incorporada no método para a escolha da variável responsável pela ramificação da árvore de classificação, ou seja, para a tomada da decisão (iii) no algoritmo 3.1.

Essa heurística foi introduzida num método de classificação designado ID3 ou *árvore induzida de decisão* (Quinlan, 1986), método esse cuja estratégia de procura de soluções e representação são semelhantes às do método proposto neste trabalho.

No método ID3 a variável escolhida para a ramificação de um nó é a que, de entre as variáveis associadas ao nó, tornar mínima a seguinte função (função de informação):

$$E(V_k) = \sum_{h=1}^{n_k} [\frac{L_{kh}}{L'} \sum_{i=1}^{m} (-\frac{L_{khi}}{L_{kh}} . log_2 \frac{L_{khi}}{L_{kh}})],$$

em que,

$$L_{khi} = card\ \{O_j : V_k(O_j) \in P_{k_h} \wedge C(O_j) = c_i\} \quad \text{e}$$

$$L_{kh} = card\ \{O_j : V_k(O_j) \in P_{k_h}\} = \sum_{i=1}^{m} L_{khi}$$

Os termos da expressão de $E(V_k)$ têm o significado apresentado em seguida.

- L' é o número de objectos associados ao nó (é a dimensão da subamostra).
- L_{kh} é o número de objectos da subamostra que têm o atributo P_{k_h}. O quociente $\frac{L_{kh}}{L}$ é, portanto, uma estimativa da probabilidade de $V_k \in P_{k_h}$ para a população representada pela subamostra.
- L_{khi} é o número de objectos da subamostra que têm o atributo P_{k_h} e que pertencem à classe c_i. Como atrás, pode afirmar-se que $\frac{L_{khi}}{L_{kh}}$ é uma estimativa de $P(c_i|V_k \in P_{k_h})$ para a população representada pela subamostra.
- $\sum_{i=1}^{m} (-\frac{L_{khi}}{L_{kh}} . log_2 \frac{L_{khi}}{L_{kh}})$ é a **entropia** [2] da classificação dos objectos da subamostra que verificam $V_k \in P_{k_h}$. O simétrico desse valor é o que se ganha, em

[2] Em classificação, e segundo Thompson *et al.* (1986), entropia é uma medida de *incerteza* sobre a classe a que pertence um objecto.

informação sobre a verdadeira classificação dos objectos da subamostra, por se saber que $V_k \in P_{k_h}$. Como $-\frac{L_{khi}}{L_{kh}} . log_2 \frac{L_{khi}}{L_{kh}}$ não está definido para $L_{khi} = 0$ convenciona-se que, nessa situação, o seu valor é nulo.

- $E(V_k)$ é o somatório, para as n_k modalidades de V_k, do termo acima e representa, finalmente, o aumento de entropia da classificação no caso de não ser conhecida o valor da variável V_k para os objectos da subamostra. E é uma função sempre não negativa que atinge o valor mínimo quando todos os objectos de cada subamostra pertencem à mesma classe e atinge o valor máximo quando estão igualmente distribuidos por todas as classes. Por esta razão E (tal como outras funções com essas características) é designada como *função de impureza* (Breiman *et al.*, 1984)

A heurística consiste, portanto, em determinar o valor de $E(V_k)$ para todas as variáveis associadas ao nó em análise e ramificar o nó segundo a variável que tiver menor $E(V_k)$. Esta regra leva a escolher a ramificação que conduz a uma maior "separação" das classes.

Poderiam ser consideradas outras heurísticas tais como: 1) estabelecer um limiar para $E(V_k)$, a partir do qual uma variável deixaria de poder ser escolhida para a ramificação (heurística integrada em CART); ou 2) testar, para cada variável disponível, a independência entre o conjunto de nós obtidos por ramificação segundo essa variável e o conjunto das classes e rejeitar, para a ramificação, as variáveis que gerem nós filhos independentes das classes (cf. Quinlan, 1986, para duas classes apenas, e Milton *et al.*, 1989, para o caso mais geral).

As duas heurísticas anteriores são aproximadamente equivalentes. Nenhuma delas será incorporada no método mas, em alternativa, será considerada uma outra heurística (v. 4.2) que, não sendo equivalente, impede também a ramificação exagerada da árvore de classificação.

4.2. Heurística para afectação de uma região de decisão a uma classe

Considere-se a tomada das decisões (i) e (ii) do algoritmo 3.1. Uma heurística para dar uma resposta a essas questões tem que, por um lado, identificar as situações em que uma nova ramificação do nó em análise não é, provavelmente, útil para a descoberta de uma boa solução (decisão i) e, por outro lado, seleccionar a classe

que deve ser afectada a esse nó (decisão ii). Frequentemente, as duas decisões estão muito relacionadas e implicam-se mutuamente.

Os critérios de paragem de ramificação num nó são os apresentados em seguida.

- O primeiro critério diz respeito a situações em que a amostra não fornece informação para tomar a decisão (ii). Essas situações são: 1) ausência de objectos no nó, isto é, não existe na amostra nenhum j tal que $V_e(O_j) \in P_{e_h}$, sendo h o índice do conjunto de valores que a variável V_e pode tomar nos objectos do nó; 2) esgotamento das variáveis, isto é, o conjunto de variáveis associadas ao nó corrente é vazio.

  Em ambos os casos, a decisão (i) consiste em não ramificar mais o nó. Em relação à decisão (ii), e na falta de outra informação (para além da incluida na amostra), o nó poderia ser afectado a uma classe artificial (classe dos objectos não classificados). Quinlan propõe, em alternativa à consideração de uma classe artificial, a afectação de cada nó não conclusivo à classe mais frequente no seu nó *pai*. Como se verá, o método proposto integra outras heurísticas que permitem que seja tomada a decisão (ii).

- O segundo critério diz respeito a situações em que os objectos da subamostra pertencem à mesma classe d, ou seja, $\forall_{O_j} : V_e(O_j) \in P_{e_h} \Rightarrow C(O_j)$ *constante*. Neste caso o nó não é ramificado e é afectado à classe a que pertencem esses objectos.

  Relaxando esta condição para atender à estocacidade dos objectos, o critério pode ser definido à custa de uma estimativa da probabilidade dos objectos do nó pertencerem a uma determinada classe. A ramificação é interrompida se a proporção dos objectos, correspondentes ao nó, da mesma classe fôr superior a um determinado valor (α, por exemplo). Seja $\mathcal{E}$ o acontecimento "pertencer ao nó". A proporção atrás referida pode ser denotada por $P_A(c_i|\mathcal{E})$, para a classe c_i (o índice A indica que a estimativa de $P(c_i|\mathcal{E})$ se baseia na amostra). Nesse caso a probabilidade de erro na classificação será, em média, sempre inferior a $(1-\alpha).p$, sendo p a probabilidade de um objecto pertencer ao nó.

4.3. Exemplo

Considere-se de novo o conjunto de dados da Tabela 2.1. e duas árvores de classificação construídas a partir desses dados que são apresentadas em seguida

(figuras 4.1 e 4.2). Como se pode verificar, as duas árvores representam regras de classificação que classificam, sem cometer nenhum erro, os objectos da amostra mas diferem na complexidade. A árvore 1 foi construída automaticamente utilizando a heurística baseada na teoria da informação e os critérios baseados na amostra para a paragem da ramificação e afectação. A árvore 1 origina cinco regiões de decisão e tem, portanto, complexidade 5.

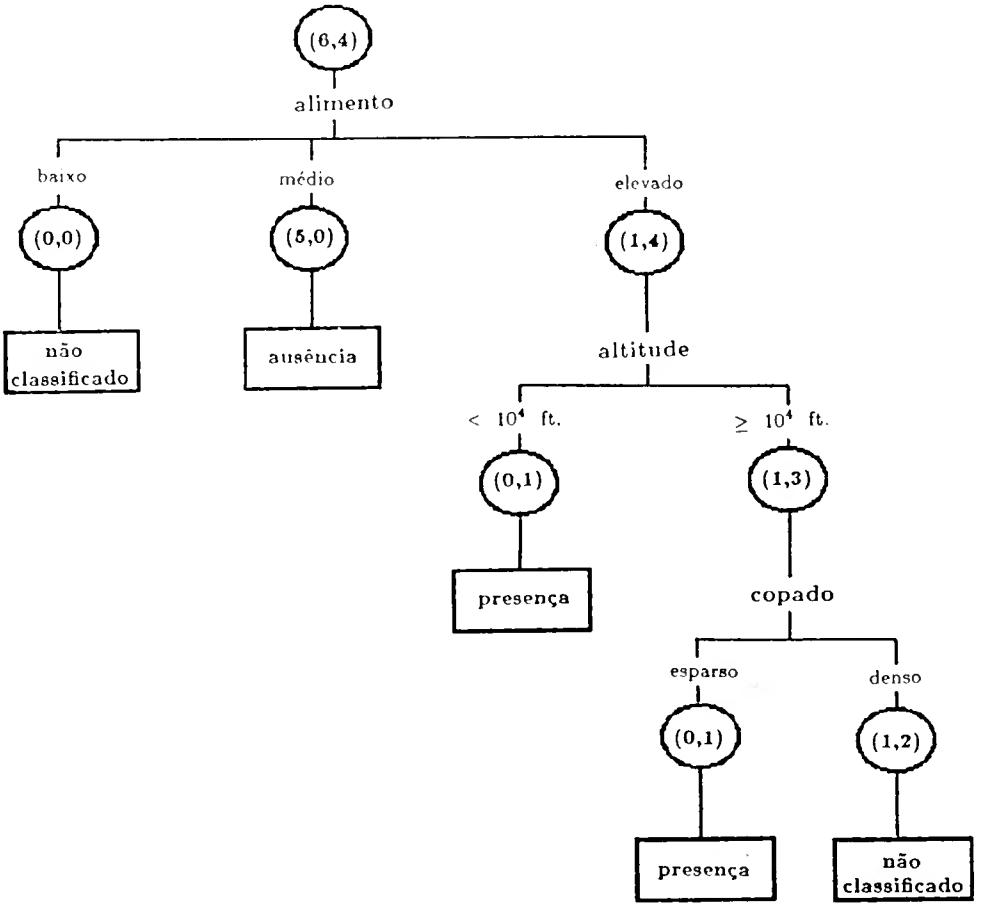


Figura 4.1 - Árvore 1, obtida pelo método ID3 [3]

A escolha da variável para a primeira ramificação da árvore 2 não foi feita segundo

[3]Os números entre parêntesis indicam o número de objectos da subamostra que pertencem, respectivamente, à classe ausência e à classe presença.

o critério da função de informação. A árvore 2 tem complexidade superior à árvore 1 pois origina mais uma região de decisão.

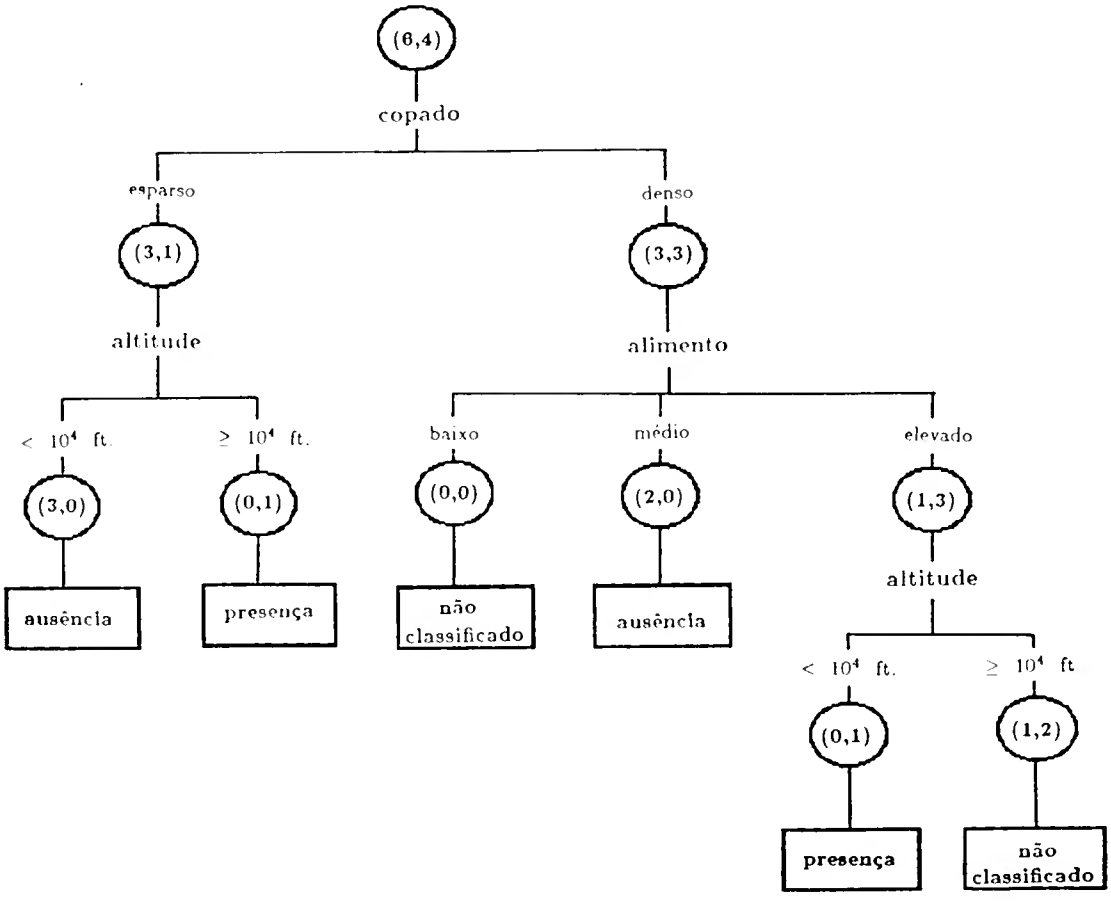


Figura 4.2. - Árvore 2.

Saliente-se que as variáveis e modalidades consideradas não permitem discriminar os objectos *parcela 2* e *parcela 6* da classe *presença* do objecto *parcela 8* da classe *ausência*.

Ambas as árvores de classificação classificam da mesma forma os objectos da amostra, mas podem classificar de forma diferente um novo objecto. Suponha-se que se pretendia classificar um objecto com atributos [*altitude* $\geq 10^4$ ft.], [*alimento=médio*] e [*copado = esparso*]. Pela árvore 1 esse objecto seria classificado na classe *ausência* e pela árvore 2 seria classificado na classe *presença*.

Considere-se, agora, uma árvore (árvore 3, apresentada na figura 4.3) também construída utilizando a heurística baseada na teoria da informação mas diferindo da árvore 1 em relação ao critério de afectação. Na árvore 3, a proporção dos objectos

nó [alimento=elevado] que são da classe *presença* é de 0.8. Considerando uma margem de erro de $(1 - \alpha) = 0.2$, foi afectado o referido nó à classe *presença*.

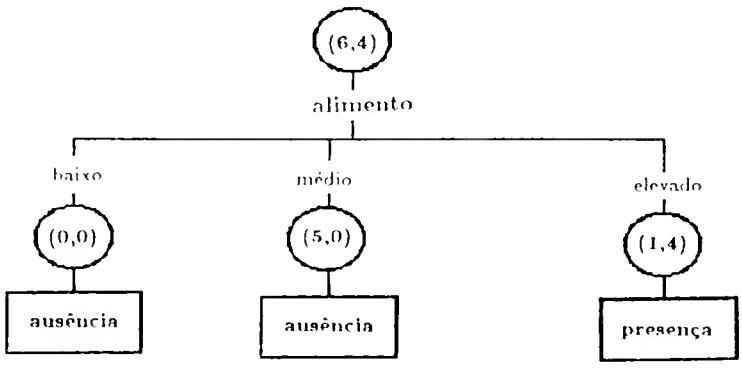


Figura 4.3 - Árvore 3

A árvore 3 distingue-se da árvore 1 nos aspectos seguintes: 1) tem menor complexidade; 2) classifica numa classe errada o objecto 8 da amostra.

Apresenta-se, apenas por curiosidade (pois a dimensão da amostra não é significativa), uma tabela (tabela 4.1) com os desempenhos das árvores 1, 2 e 3 na classificação dos 471 objectos que constituem a população considerada na aplicação real a apresentar no capítulo 7.

	% de objectos		
árvores	bem classif.	não classif.	mal classif.
1.	45.4	33.1	21.5
2.	54.0	26.0	20.0
3.	74.7	-	25.3

Tabela 4.1 - Taxas de erro reais para as árvores 1, 2 e 3

Ainda sobre este exemplo importa realçar o seguinte aspecto. Nas árvores 1 e 3 a asserção [*alimento=baixo*] é afectada a uma classe artificial embora a asserção [*alimento=médio*] seja afectada à classe *ausência*. Isso não parece fazer sentido

pois não é natural que não se possa prever se o animal (cuja presença ou ausência define as classes) está presente em parcelas com o atributo [*alimento=baixo*] e que se preveja que não está presente em parcelas com o atributo [*alimento=médio*]. Este aparente "erro" da regra de classificação (que poderia, provavelmente, não se verificar ao nível das primeiras ramificações da árvore se a amostra fosse de dimensão significativa porque a probabilidade de, pelo menos, um objecto da amostra pertencer ao nó cresce com a dimensão da amostra) chama a atenção para as vantagens da incorporação, no método de classificação, de heurísticas baseadas no conhecimento pericial sobre o domínio do problema.

5 . Heurísticas baseadas num modelo de representação do conhecimento pericial sobre o domínio do problema

O método de classificação proposto neste trabalho envolve um modelo de relações entre as variáveis envolvidas no processo de classificação. Com base nesse modelo, que integra um conjunto de regras de dependência entre grupos de variáveis, são definidas heurísticas para auxiliar a procura da solução do problema de classificação.

O modelo é, numa primeira fase do processo de classificação, construido através de um processo interactivo entre um perito e o programa. Nessa fase, a informação do perito é confrontada com a informação da amostra e é incorporada no modelo. Numa segunda fase, o modelo é utilizado, de forma heurística, para orientar o método de classificação.

Em 5.1 será descrito o modelo de representação do conhecimento (um grafo bayesiano). Em 5.2 e 5.3 serão caracterizadas as heurísticas construídas com base nesse modelo e em 5.4 serão exemplificadas as noções apresentadas.

5.1. Grafo bayesiano

O grafo bayesiano, também designado *belief network*, *influence network* (Pearl, 1987) ou *probabilistic influence diagram* (Geiger *et al.*, 1990), é, formalmente, um grafo orientado acíclico construído a partir de uma distribuição de probabilidade de um conjunto de variáveis aleatórias e será representado por $\mathcal{G} = (G, p)$, em que G representa o grafo orientado acíclico e p a distribuição de probabilidade.

O grafo bayesiano é constituido pelas duas seguintes componentes. A primeira designa-se componente gráfica e consiste num modelo gráfico de relações de dependência entre variáveis. Este modelo representa o conhecimento *qualitativo* sobre o domínio do problema pois apenas exprime a existência de relações de dependência entre as variáveis envolvidas, sem associar uma medida de grandeza a essas relações. A segunda componente designa-se componente probabilística e caracteriza quantitativamente as relações expressas pelo modelo gráfico. Designa-se probabilística porque se baseia na teoria das probabilidades para associar medidas de incerteza ao modelo de relações entre as variáveis.

5.1.1. Componente gráfica

5.1.1.1. Definições prévias [1]

Um grafo $G = (V, E)$ é constituido por um conjunto de vértices ($V = \{V_1, ..., V_p\}$) e por um conjunto de arestas ($E = \{E_1, ..., E_q\}$). Cada aresta é definida por um par de vértices $E_k = (V_i, V_j)$, que pode ser ordenado (neste caso a aresta diz-se orientada e pode ser representada por $V_i \rightarrow V_j$) ou não ordenado (que corresponde, portanto, a uma aresta não orientada e pode ser representada por $V_i \leftrightarrow V_j$). Um grafo é designado por **grafo orientado** se todas as suas arestas forem orientadas e por **grafo não orientado** se todas as suas arestas forem não orientadas. Define-se **grafo não orientado associado a um grafo orientado** como sendo o grafo $G' = (V', E')$ associado a $G = (V, E)$ orientado, com $V' = V$ e sendo os elementos de E' as arestas não orientadas definidas pelos mesmos pares de vértices que definem os elementos análogos de E. Um **caminho** entre os vértices V_i e V_j é uma sequência de vértices distintos $V_i = \alpha_0, \alpha_1, ..., \alpha_n = V_j$ tais que $\alpha_{i-1} \rightarrow \alpha_i$ (num grafo orientado) ou $\alpha_{i-1} \leftrightarrow \alpha_i$ (num grafo não orientado) para todo o $i = 1, ..., n$. Uma **cadeia** é uma sequência de vértices distintos de um grafo orientado $V_i = \alpha_0, \alpha_1, ..., \alpha_n = V_j$ tais que $\alpha_{i-1} \rightarrow \alpha_i$ ou $\alpha_i \rightarrow \alpha_{i-1}$ para todo o $i = 1, ..., n$. São, portanto, iguais, as cadeias de um grafo orientado e os caminhos do grafo não orientado associado. Um grafo G orientado sem ciclos será representado por DAG G. Um grafo é **conexo** se existir pelo menos um caminho entre qualquer par dos seus vértices. Um grafo diz-se **completo** se existir uma aresta entre qualquer par de vértices. Uma **clique** é um subgrafo completo e maximal. Num grafo orientado, se $V_i \rightarrow V_j$ diz-se que V_i é *pai* de V_j e que V_j é *filho* de V_i. Os conjuntos $pa(V_i)$ e $fi(V_i)$ são, respectivamente, os conjuntos dos *pais* e dos *filhos* de V_i. Evidentemente, qualquer um desses conjuntos pode ser vazio. Denota-se por $pa_j(V_i)$ o j-ésimo *pai* do vértice V_i e por npa_i o número de *pais* de V_i. Um vértice que não tem nenhum *filho* designa-se **vértice terminal**. Os *descendentes* de V_i, $dec(V_i)$, são os vértices de um DAG G para os quais existe, pelo menos, um caminho entre V_i e esses vértices. De forma análoga se define $asc(V_i)$ como o conjunto dos *ascendentes* de V_i. Qualquer que seja $\alpha \in V$ de um DAG G, $asc(\alpha)$ e $dec(\alpha)$ são disjuntos. Um subconjunto $A \subset V$ diz-se **ancestral** se $asc(\alpha) \subset A$ para todo o $\alpha \in A$.

Uma noção importante é a de separador. Pearl (1986) definiu da seguinte forma o que viria a ser designado posteriormente, por este e outros autores,

[1] Algumas definições já foram apresentadas em 3.1.1. mas são apresentadas novamente por forma a se distinguirem bem as diferenças entre noções aplicáveis a grafos orientados e não orientados e por forma a haver uma concordância com a notação utilizada neste capítulo.

d-separador entre os vértices α e β num grafo orientado:
(a) $S \subset V$ d-separa α de β se todas as cadeias entre α e β são separadas por S;
(b) Uma cadeia P é separada por S se pelo menos três sucessivos vértices de P estão *bloqueados* por S;
(c) Três vértices sucessivos (α, β, γ) de uma cadeia estão bloqueados por S se $\beta \in S$ e se $\alpha \rightarrow \beta \rightarrow \gamma$ ou $\alpha \leftarrow \beta \leftarrow \gamma$ ou $\alpha \leftarrow \beta \rightarrow \gamma$;
(d) Três vértices sucessivos (α, β, γ) de uma cadeia estão bloqueados por S se $\alpha \rightarrow \beta \leftarrow \gamma$ e se nem β nem nenhum dos seus descendentes pertencer a S.
O termo **separador** é utilizado para grafos não orientados. Segundo Lauritzen *et al.* (1990), um subconjunto $S \subset V$ separa $A \subset V$ de $B \subset V$ se todos os caminhos entre $\alpha \in A$ e $\beta \in B$ contém pelo menos um elemento de S.

5.1.1.2. O modelo gráfico

O modelo gráfico representa a estrutura qualitativa do problema. Os vértices representam as variáveis envolvidas e cada aresta a existência de uma relação relevante entre um par de variáveis.

Neste trabalho G denotará o grafo constituído pelo conjunto de vértices $V^G = \{V_1, ..., V_p, V_{p+1}\}$, em que $\{V_1, ..., V_p\}$ é o conjunto de variáveis que descrevem os objectos a classificar e V_{p+1} é uma variável aleatória cujo domínio é o conjunto de classes definidas à partida às quais podem pertencer esses objectos, e pelo conjunto de arestas orientadas que representam, cada uma, a existencia de uma relação *directa* entre um par de ariáveis de V^G. G pode, com rigor, ser considerado um hipergrafo porque a representação da relação existente entre um vértice e os seus *pais* (se tiver mais do que um único *pai*) requere uma função com mais do que dois argumentos (a própria variável e os seus *pais*). No entanto, esse facto não retira as vantagens do grafo como estrutura de representação desde que se tenha sempre presente que cada variável V_i depende directamente do conjunto $pa(V_i)$ e não de apenas algum(uns) dos seus elementos.

O grafo bayesiano é uma estrutura tem sido alvo de grande interesse pois permite representar de uma forma muito atraente uma estrutura fundamental do conhecimento humano que se baseia no estabelecimento de um conjunto de proposições (representadas pelos vértices do grafo) que descrevem a realidade e na pesquisa e actualização do conhecimento por sequências de inferências sobre as relações relevantes entre essas proposições (relações que são representadas pelas arestas do grafo). O grafo bayesiano pode ser visto não apenas como uma estrutura de representação do conhecimento mas também como uma arquitectura computacional para raciocinar

sobre esse conhecimento (Pearl, 1986). Na verdade, como se verá em 5.3, o modelo de propagação de uma informação pelo grafo bayesiano é constituido por um conjunto de processadores que realizam inferências individuais. O efeito do conhecimento de uma informação em todo o grafo é obtido através de uma sequência dessas inferências. No pequeno exemplo que será apresentado em 5.4 essa arquitectura será ilustrada. O grafo bayesiano é um exemplo, bem estudado, de sistema de produção (v. 3.4).

5.1.2. Componente probabilística

O modelo de incerteza associado à estrutura de representação do conhecimento aqui considerada - o grafo bayesiano - irá ser explorado com duas finalidades. A primeira envolve a resposta à seguinte questão: *O conhecimento do valor de uma determinada variável, sendo conhecida uma determinada evidência, tem ou não interesse para prever o valor de outra?*. A segunda diz respeito à possibilidade de estimar a distribuição de probabilidade, para cada uma das variáveis não instanciadas, dada uma evidência.

5.1.2.1. Definições prévias

Considerem-se as variáveis aleatórias discretas $V_1, ..., V_{p+1}$ e sejam $v_1, ..., v_{p+1}$ valores que essas variáveis podem tomar. Suponha-se que cada variável pode tomar um número finito de valores, isto é, $v_i \in \{v_{i_1}, ..., v_{i_{n_i}}\}$ e represente-se por $P(v_i)$ a probabilidade $P(V_i = v_i)$ e por $p(v_i)$ a distribuição de probabilidade de V_i. $p(v_i)$ também pode ser designada como distribuição de probabilidade marginal de V_i se se considerar que existe uma distribuição conjunta para $v_1, ..., v_{p+1}$. Seja $p(V) = p(v_1, ..., v_{p+1})$ a distribuição de probabilidade conjunta e $p(v_i|A^*)$, com $A \subset V$, a distribuição de probabilidade de V_i condicionada por uma instanciação das variáveis de A, A^*.

Sejam X, Y, Z variáveis aleatórias. Diz-se que X é **condicionalmente independente** de Y dado Z (o que se representa por $X \perp\!\!\!\perp Y|Z$) sse $p(x|y,z) = p(x|z)$ (Dawid, 1979).

Por **evidência** designa-se uma determinada instanciação de um subconjunto de variáveis de V, ou seja, e denotando através de $\mathcal{E}$ uma evidência, $\mathcal{E} = [V_{e_1} = v_{e_1}] \wedge ... \wedge [V_{e_{n_e}} = v_{e_{n_e}}]$. De acordo com a terminologia definida no capítulo 2, $\mathcal{E}$ é uma asserção.

5.1.2.2. Independência condicional

Suponha-se que $V = \{V_1, ..., V_{p+1}\}$ constitui o conjunto de vértices de um grafo bayesiano e que $S \subset V$. Então, é verdadeira (ver a demonstração em Lauritzen *et al:* , 1990) a seguinte afirmação:

proposição 5.1 Se V_i e V_j estão d-separados por S então $V_i \perp\!\!\!\perp V_j|S$

Este resultado permite responder à questão colocada acima. Fornece um critério, para o problema de decisão a que este trabalho diz respeito, que garante que a informação sobre V_i é desnecessária para prever V_j desde que S seja conhecido. Evidentemente esta *regra* só é válida se o grafo representar correctamente as relações entre as variáveis.

5.1.2.3. Propagação de evidências

Considere-se, agora, o segundo aspecto referido. Em particular, e porque o problema em estudo é um problema de classificação, é muito interessante determinar a distribuição de probabilidade de V_{p+1} dada uma determinada evidência.

Para tal é necessário *propagar* essa evidência pelo grafo. Serão, em seguida, apresentados com maior ou menor detalhe alguns métodos para realizar esse objectivo, sendo, finalmente, feita a sua comparação de acordo com o quadro geral deste trabalho.

5.1.2.3.1. Métodos de propagação

Todos os métodos que serão aqui descritos baseam-se no seguinte resultado que permite pôr em prática o princípio da computação local:

proposição 5.2 A distribuição conjunta das variáveis $V_1, ..., V_{p+1}$ é igual ao produto das distribuições de probabilidade de cada uma condicionadas pelos seus ***pais***, isto é:

$$p(V) = \prod_{i=1}^{p+1} p(v_i|pa(v_i))$$

demonstração: A distribuição conjunta dessas variáveis pode ser considerada (v. Murteira, 1979, teorema 1.12) como o produto de probabilidades condicionadas $P(v_1, ..., v_{p+1}) = P(v_1)P(v_2|v_1)P(v_3|v_1, v_2)...P(v_{p+1}|v_1, ..., v_p)$.
Se a sequência $V_1, ..., V_{p+1}$ for tal que $asc(V_i) \subset \{V_1, ..., V_{i-1}\}$, $\forall_{i \in \{1,...,p+1\}}$, o que é sempre possível por G ser orientado e acíclico, então garantimos que $pa(V_i) \subset \{V_1, ..., V_{i-1}\}$.
Fazendo $av(V_i) = \{V_1, ..., V_{i-1}\} - pa(V_i)$ e sabendo que, pelas propriedades de

$\mathcal{G}$, $V_i \perp\!\!\!\perp av(V_i)|pa(V_i)^*$, vem, pela definição de independência condicionada atrás apresentada, que $P(v_i|av(V_i)^*, pa(v_i)^*) = P(v_i|pa(v_i)^*)$, $\forall_i$, provando-se assim o resultado. •

Saliente-se a importância, do ponto de vista operacional, da proposição 5.2 para a estimação das distribuições marginais das variáveis. A estimação directa de $p(V)$ a partir da amostra, não considerando o modelo gráfico, exige uma enorme quantidade de dados e torna-se, na prática, impossível. Na hipótese do modelo gráfico ser correcto, a estimação de $p(V)$ reduz-se à estimação das distribuições das probabilidades condicionadas $p(v_i|pa(v_i))$. Esta forma de estimar $p(V)$ envolve, geralmente, um número muito menor de parâmetros.

A igualdade $p(V) = \prod_{i=1}^{p+1} p(v_i|pa(v_i))$ permite decompor a distribuição de probabilidade para todo o grafo num conjunto de distribuições que envolvem apenas subconjuntos restritos das variáveis e permite, portanto, construir um algoritmo de propagação recursivo sobre os subconjuntos $(\{V_i\} \cup pa(V_i))$.

Veja-se, agora, o que se pode afirmar sobre a distribuição de probabilidade de uma variável do grafo, no caso em que as restantes variáveis têm um valor fixo.

proposição 5.3 A distribuição de probabilidade de V_i, $\forall_{i\in\{1,\ldots,p+1\}}$, condicionada pelo estado de todas as outras variáveis, é dada pelo produto [2]

$$P(v_i|V^* - \{v_i\}) = \textit{constante} \; . \; P(v_i|pa(V_i)^*) \prod_{V_j \in fi(V_i)} P(v_j^*|pa(V_j)^*)$$

A demonstração é dada por Pearl (1987) e baseia-se no facto de, no caso de todas as variáveis com excepção de V_i terem um valor conhecido, os termos do produto da proposição 5.2 que não constam da equação apresentada na proposição 5.3 serem constantes.

O mesmo autor, com base neste resultado, propôs um algoritmo para propagação de uma evidência por simulação estocástica, ou seja, através de um método de cálculo de probabilidades por contagens das frequências dos acontecimentos numa série de simulações (**método 1**, que abaixo será descrito). Este método é adequado para aplicações que envolvem modelos complexos, mesmo com variáveis muito interdependentes (pois a complexidade computacional do algoritmo é polinomial no número total de vértices e de arestas (Pearl, 1987), não dependendo do tipo de dependências existentes), e em relação às quais é suficiente obter valores aproxima-

[2] como $V_i \in pa(V_j)$, $pa(V_j)^*$ representa neste caso $(v_i, v_\alpha^*, v_\beta^*, \ldots)$, em que $v_\alpha^*, v_\beta^*, \ldots$ são os estados dos outros *pais* dos *filhos* de V_i

dos para as distribuições de probabilidade. O grau de precisão dessas estimativas é função do número de iterações.

Os dados necessários para a execução deste algoritmo são:

- as relações de dependência entre as variáveis (relações representadas pelo grafo),
- um estado para as variáveis fixadas à partida ($\mathcal{E}$) e um estado inicial para as outras variáveis,
- o conjunto das distribuições de probabilidade condicionadas $p(v_i|pa(V_i)^*)$, $\forall_{pa(V_i)^*}$, $\forall_{i\in\{1,...,p+1\}}$.

O algoritmo trata, em cada iteração, todas as variáveis sucessivamente, de acordo com uma ordem consistente com o grafo bayesiano, isto é, $V_1, ..., V_{p+1}$ tal que $asc(V_i) \subset \{V_1, ..., V_{i-1}\}$, $\forall_{i\in\{1,...,p+1\}}$. Após cada iteração fica determinado um novo estado corrente ($\mathcal{E}$) para as variáveis que não estão fixadas em $\mathcal{E}$. Os cálculos realizados em relação a cada variável não instanciada ($V_i \notin \{V_{e_1}, ..., V_{e_{n_e}}\}$) incluem os seguintes passos:

passo 1. Determinação de $P(v_i|\mathcal{E}, \mathcal{E} - v_i)$, $v_i = v_{i_1}, ..., v_{i_{n_i}}$ através da expressão da proposição 5.3. Essas probabilidades são calculadas localmente pois dependem apenas dos valores dos *pais*, dos *pais* dos *filhos* de V_i e das distribuições de probabilidade condicionadas para V_i e $V_j \in fi(V_i)$, dados os valores (estados correntes) dos respectivos pais.

passo 2. Actualização do estado de V_i. É escolhido aleatoriamente um novo valor em $\{v_{i_1}, ..., v_{i_{n_i}}\}$ de acordo com as probabilidades calculadas no passo anterior.

Evidentemente, quanto maior for o número de simulações maior será a possibilidade de ser escolhido pelo menos uma vez um estado ao qual corresponde uma probabilidade condicionada baixa. Pearl (1987) verificou, numa aplicação que envolvia um grafo com cinco vértices, que era necessária uma centena de iterações para obter uma estimativa de $P(v_i|\mathcal{E})$ com um erro inferior a 1%.

A estimativa da distribuição de probabilidade marginal para cada variável não instanciada é obtida calculando a média das probabilidades condicionadas obtidas nas várias iterações para os diversos valores possíveis, isto é, $P(v_i|\mathcal{E}) = \frac{1}{n} \sum_{k=1}^{n} P(v_i|\mathcal{E}, \mathcal{E}(k) - \{v_i\})$, $v_i = v_{i_1}, ..., v_{i_{n_i}}$, sendo n o número de iterações e $\mathcal{E}(k)$ o estado, na iteração k, das variáveis não fixadas em $\mathcal{E}$.

Em alternativa ao método posposto em (Pearl, 1987) existem métodos de propagação de evidências exactos e igualmente baseados no paradigma da computação local, isto é, utilizando o grafo como uma arquitectura computacional

representável por um conjunto de processadores autónomos. Se esses processadores (também designados *belief universes*) forem de pequena dimensão então a complexidade computacional dos cálculos necessários para a propagação de uma evidência pelo grafo inicial, ao serem substituídos por um conjunto de cálculos em cada um dos processadores, será consideravelmente diminuída, ultrapassando-se desta forma o problema da dimensionalidade.

Spiegelhalter *et al.* (1992) (**método 2**) apresentam um quadro formal que permite substituir o DAG G inicial por uma estrutura modular mantendo as relações de dependência entre variáveis. Essa estrutura é uma árvore construída da seguinte forma:

passo 1. Adicionar arestas a G de forma a que $\forall_{V_i}$, $pa(V_i) \cup \{V_i\}$ forme um subgrafo completo.

passo 2. Construir o grafo não orientado associado (que é denotado por G^m - *moral graph* - pois todos os pares de *pais* do mesmo vértice estão unidos por uma aresta em G^m).

passo 3. Adicionar arestas de forma a que G^m se transforme num grafo triangulado [3] (G^T).

passo 4. Construir uma árvore $\mathcal{T}$ (*junction tree*) cujos vértices são as cliques de G^T e cujas arestas ligam os pares de cliques que têm elementos comuns.

Este processo merece algumas considerações. O passo 3. é sempre possível mas apenas é útil, pelo que foi referido atrás, se originar subgrafos completos de pequena dimensão. Ora, esse problema de optimização é NP-completo (Spieghelhalter, 1992), o que constitui a principal limitação deste método. No que respeita à noção de independência condicional, o modo de construção de G^m fornece um critério (demonstrado em Lauritzen *et al.*, 1990, proposição 3.) mais simples do que o apresentado na proposição 5.1:

proposição 5.4 Sejam A, B, S subconjuntos disjuntos de V. Se S separa A de B em $G^m_{asc(A \cup B \cup S)}$ [4] então $A \perp\!\!\!\perp B|S$.

Em relação à propagação de uma evidência, esta pode ser realizada localmente em cada vértice de $\mathcal{T}$, sendo os vértices processados sequencialmente de acordo com a estrutura de $\mathcal{T}$.

[3] Um grafo triangulado é um grafo não orientado e cujos ciclos não têm mais do que três arestas.
[4] $G^m_{asc(A \cup B \cup S)}$ é o *moral graph* do subgrafo de DAG G constituido pelos ascendentes de A, B e S.

Pearl (1986) apresenta uma abordagem diferente para o problema (**método 3**). O método descrito por este autor passa pela transformação do grafo numa árvore através da utilização de variáveis auxiliares. Essa transformação apenas é possível se se verificarem algumas condições restritivas, nomeadamente: as variáveis devem ser binárias, a árvore deve, por hipótese, existir e os valores das correlações entre vértices terminais devem ser conhecidos. No mesmo artigo o autor apresenta um método eficiente de propagar uma evidência através de uma árvore.

5.1.2.3.2. Escolha do método de propagação

Pretende-se, para atingir os objectivos deste trabalho, dispôr de uma estrutura de armazenamento do conhecimento fléxivel, isto é, uma estrutura que possa ser alterada no decorrer do processo de classificação. Nesse sentido é necessário executar um algoritmo de propagação directamente sobre G e não sobre uma estrutura obtida, com elevado custo computacional, a partir de G. Spiegelhalter *et al.* (1992) reconhecem que, no método 2, essa transformação pode ser computacionalmente exigente embora argumentem, no quadro do seu trabalho, que só é necessário realizá-la uma única vez. Caso sejam ultrapassados os problemas computacionais, o método 2 é sempre aplicável mas pode ser muito pouco interessante se G for tal que não seja possível construir grafo triangulado (G^T) cujas cliques sejam de pequena dimensão (esta situação é frequente no caso das variáveis envolvidas serem razoavelmente interdependentes). O método 3, para além de se aplicar apenas a grafos bayesianos cujos vértices sejam variáveis aleatórias binárias, não garante a existência de uma solução. O método 1 tem como principal desvantagem o facto de dar valores aproximados para as probabilidades marginais. Essa desvantagem pode ser atenuada aumentando o número de iterações e não afecta sensivelmente a qualidade do método de classificação proposto neste trabalho pois o objectivo é captar o conhecimento proveniente do perito e da amostra que é, forçosamente, aproximado (o perito poderá supôr uma determinada distribuição de probabilidade mas associa-lhe uma margem de erro que, geralmente, não será inferior a 5% para cada probabilidade). Devido ao conjunto de razões referidas, o método 1, de propagação de evidências pelo grafo, será adoptado.

5.1.2.4. Considerações adicionais sobre a informação necessária

Importa tecer algumas breves considerações sobre os dados necessários para a execução de qualquer um dos métodos descritos e, em particular, sobre a distribuições de probabilidade das variáveis condicionadas por todos os estados possíveis dos seus *pais*.

Por um lado, definição dessas distribuições exige bastante informação, obtenível através da amostra e através do conhecimento do perito. A quantidade de valores necessários para definir o modelo de incerteza pode ser muito elevada [5] e, portanto, é conveniente evitar que o perito tenha que definir a totalidade desses valores. Tirando partido da amostra é possível ultrapassar esse problema e, simultaneamente, controlar o conhecimento pericial. É necessário, para realizar a propagação de uma evidência, conhecer os valores $P(v_i|pa(V_i)^*)$ para todas as modalidades de todas as variáveis e para todos os estados possíveis dos seus *pais*. Pela definição de probabilidade condicionada,

$$P(v_i|pa(V_i)^*) = \frac{P(v_i, pa(V_i)^*)}{P(pa(V_i)^*)} \quad \text{e} \quad \sum_{v_i} P(v_i|pa(V_i)^*) = 1.0,$$

e portanto, é apenas necessário conhecer os valores de $P(v_i, pa(V_i)^*)$ pois $P(pa(V_i)^*)$, que é independente das modalidades de V_i, não é mais do que uma constante de normalização das probabilidades. Uma estimação trivial, com base na amostra, dos valores de $P(v_i, pa(V_i)^*)$ para todos os v_i, todos os i e todos os $pa(V_i)^*$ associados é dada pela proporção dos objectos da amostra que pertencem à extensão da asserção $[V_i = v_i] \wedge [pa_1(V_i) = pa_1(V_i)^*] \wedge ... \wedge [pa_{npa_i}(V_i) = pa_{npa_i}(V_i)^*]$. No entanto, e salvo o caso em que a dimensão da amostra seja muito grande, uma grande parte dessas estimativas é constituída por valores nulos. É possível evitar essa situação utilizando outro método de estimação que se baseia numa função que *aplana* a distribuição de frequências como seja o estimador por núcleo de Rosenblatt (Tomassone *et al.*, 1988). Esse método mais complexo de estimação não é considerado neste trabalho porque, em alternativa, é utilizada uma técnica simples que consiste em substituir os valores nulos das estimativas das probabilidade por valores baixos (0.02, por exemplo). Como os valores iniciais (estimados) são revistos pelo perito na fase de captação do conhecimento pericial, a técnica utilizada é suficiente para "aplanar", numa primeira aproximação, as distribuições de probabilidades condicionadas.

Existindo uma estimativa das probabilidades condicionadas necessárias, é fácil incorporar o conhecimento pericial. Para tal, basta que o perito modifique os valores das probabilidades que considera inaceitáveis em função do seu conhecimento sobre as relações entre as variáveis.

Seria possível, por outro lado, considerar a variabilidade dos parâmetros das distribuições de probabilidade. Spiegelhalter *et al.* (1990) propõem para esse fim uma abordagem estatística baseada na distribuição de Dirichlet ou no modelo logístico.

[5] São necessários $\sum_{i=1}^{p}(n_{V_i} \cdot \prod_j n_{pa_j(V_i)})$ valores, sendo n_{V_i} o número de modalidades da variável V_i e $n_{pa_j(V_i)}$ o número de modalidades do j-ésimo *pai* de V_i.

5.2. Heurística para a escolha da ramificação da árvore de classificação

Como no caso da heurística baseada na amostra, é possível retirar do grafo bayesiano uma regra para auxiliar a escolha da variável responsável pela ramificação do nó da árvore de classificação em análise, ou seja, uma regra para a tomada da decisão (iii) no algoritmo 3.1.

A heurística consiste em não escolher, como variável responsável pela ramificação, uma variável condicionalmente independente, dado o grafo bayesiano, do conjunto de variáveis que definem a asserção correspondente ao nó em análise. Esta heurística não permite, em geral, escolher uma determinada variável mas, somente, excluir um subconjunto de variáveis do conjunto das que estão associadas ao nó.

5.3. Heurística para afectação de uma região de decisão a uma classe

Considere-se, mais uma vez, o algoritmo 3.1 e as tomadas de decisão (i) e (ii). O grafo bayesiano fornece, como foi atrás descrito, um modo de estimar a distribuição marginal de todas as variáveis do grafo não fixadas numa determinada evidência, ou seja, todas as variáveis que não definem a asserção correspondente ao nó em análise. Em particular, fornece uma estimativa da distribuição marginal da variável classe. Os valores que compõem essa distribuição podem ser representados por $P(c_i|\mathcal{E},\mathcal{G})$, para todas as classes $c_i \in \mathcal{C}$, pois são valores das probabilidades de ocorrência das classes c_i, dada uma determinada evidência $\mathcal{E}$ e uma determinada configuração do grafo bayesiano, $\mathcal{G}$.

Importa salientar que, embora o formalismo de representação do conhecimento conduza à obtenção de entidades semelhentes (probabilidades), as heurísticas baseadas na amostra e as baseadas no conhecimento pericial têm uma origem distinta. Os valores estimados para $P(c_i|\mathcal{E},\mathcal{G})$ são calculados (v. 5.1) com base num modelo do conhecimento pericial sobre o domínio do problema (o grafo bayesiano). No método de classificação proposto, a informação da amostra é utilizada pelo perito para controlar as suas hipóteses sobre as relações entre as variáveis envolvidas no problema e, portanto, afecta indirectamente os valores estimados para $P(c_i|\mathcal{E},\mathcal{G})$. No entanto, esses valores poderiam ser obtidos mesmo que não existisse nenhuma amostra de objectos de classificação conhecida.

Do grafo bayesiano é possível construir uma regra baseada nos valores de

probabilidade atrás referidos para a tomada das decisões (i) e (ii) no algoritmo 3.1. A decisão (i) consiste em não ramificar o nó corrente se algum dos valores $P(c_i|\mathcal{E},\mathcal{G})$ fôr superior a α, sendo então $(1-\alpha).P(\mathcal{E})$ a probabilidade esperada máxima de erro na classificação. A decisão (ii) consiste em afectar o nó à classe que torna máximo $P(c_i|\mathcal{E},\mathcal{G})$.

A forma concreta de utilização destas heurísticas no método de classificação será descrita no capítulo seguinte, no qual será explicado o modo de integração de todas as heurísticas consideradas.

5.4. Exemplo

Retomando o pequeno problema que tem sido utilizado nos capítulos anteriores para ilustrar as noções apresentadas, irão ser exemplificadas, em seguida, as noções de componente gráfica do grafo bayesiano, independência condicional (para este caso será o problema será ligeiramente alterado de modo a ser considerada mais uma variável), componente probabilística e propagação de uma evidência pelo grafo segundo o método 1.

Considere-se, mais uma vez, a amostra descrita na tabela 2.1.. Três variáveis caracterizam os objectos (*altitude, alimento e copado*). Um perito sobre o domínio em questão poderia supôr a existência de relações entre essas três variáveis e a classe. Considerem-se as hipóteses referidas em 2.6: 1) as condições de desenvolvimento das copas são mais favoraveis em zonas de altitude elevada; 2) a densidade do copado favorece a produção de alimento; e 3) boas condições de copado e de alimento favorecem a presença do animal.

Estas hipóteses são facilmente representadas pelo modelo gráfico da figura 5.1.

Como se observa nessa figura, o modelo é claro e facilmente interpretável.

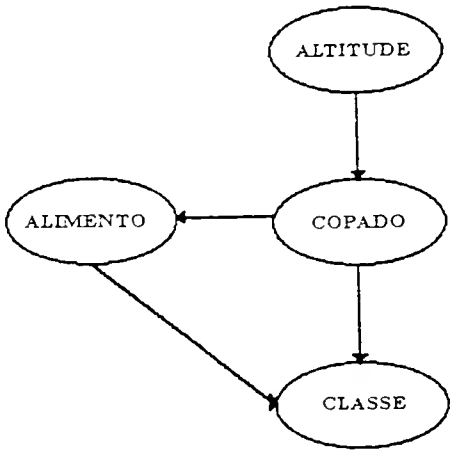


Figura 5.1 - Modelo gráfico que representa as hipóteses do perito sobre o problema apresentado em 2.6

Por forma a ilustrar a noção de separador é necessário observar um grafo ligeiramente mais complexo. Suponha-se que, para o mesmo problema, os objectos (parcelas de terreno) são também caracterizados pela menor distância que os separa de zonas de clareira (zona sem vegetação). A variável que regista esse valor é a variável *distância* e tem duas modalidades ([$<$ 5 *parcelas*] e [$\geq$ 5 *parcelas*]). O perito põe, sobre o novo conjunto de variáveis, duas hipóteses adicionais:

- a distância é baixa predominantemente nas zonas elevadas (esta hipótese deve-se ao facto do perito saber que as estradas, que são consideradas zonas de clareira, localizam-se nos cumes);
- o animal prefere locais que estejam afastados das clareiras.

O modelo gráfico que descreve as hipóteses sobre as variáveis agora consi-

deradas consta da figura 5.2.

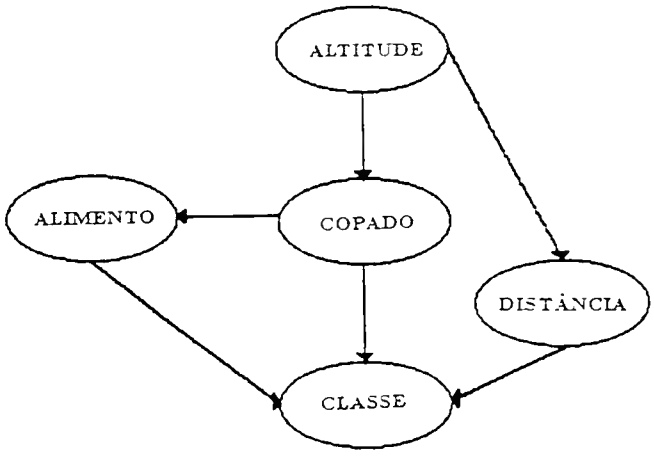


Figura 5.2 - Modelo gráfico considerando a variável *distância*

Pode afirmar-se que $\{copado\}$ é um d-separador de $\{altitude\}$ e $\{alimento\}$ pois as quatro cadeias existentes entre os vértices *altitude* e *alimento* estão bloqueados por *copado*. A cadeia $altitude \rightarrow copado \rightarrow alimento$ e a cadeia $altitude \rightarrow distância \rightarrow classe \leftarrow copado \rightarrow alimento$ têm um bloqueio do tipo (c) (v. 5.1.1.1), a cadeia $altitude \rightarrow distância \rightarrow classe \leftarrow alimento$ tem um bloqueio do tipo (d) e a cadeia $altitude \rightarrow copado \rightarrow classe \leftarrow alimento$ está bloqueada das duas formas. Pela proposição 5.1 pode afirmar-se que o conhecimento do valor de alimento não é útil para prever o valor de altitude e vice-versa, dado um valor conhecido para *copado* e na hipótese do modelo gráfico estar correcto.

Observe-se, ainda, a figura 5.2. Suponha-se, agora, que são conhecidos os valores das variáveis *copado* e *classe* (no caso anterior supunha-se que era conhecido apenas o valor de *copado*). Neste caso a cadeia $altitude \rightarrow distância \rightarrow classe \leftarrow alimento$ não está bloqueada por $S = \{copado, classe\}$ e, portanto, não se pode afirmar que *alimento* é condicionalmente independente de *altitude* dado S.

Exemplifique-se, em seguida, a noção de propagação de uma evidência pelo grafo bayesiano. Observe-se, novamente, a figura 5.1 e tente-se prever o efeito de uma evidência (por exemplo, $\mathcal{E}=[altitude < 10^4 \text{ ft.}] \wedge [alimento = elevado]$) sobre as restantes variáveis.

Pela proposição 5.2, pode-se escrever [6]

$$p(V) = p(alt, cop, ali, cla) = p(alt).p(cop|alt).p(ali|cop).p(cla|cop, ali)$$

Desde que sejam conhecidas as distribuições condicionadas envolvidas é possível determinar $p(V)$ e, a partir dessa distribuição, determinar as distribuições marginais das variáveis.

A propagação de $\mathcal{E}$ pelo método de Pearl seria executada seguindo os seguintes passos.

inicialização: escolher um estado inicial para todas as variáveis não fixadas (por exemplo, $[cop = denso]$ e $[cla$=*presença*$]$) e escolher uma ordem conveniente para o processamento dos vértices (essa ordem deve ser $\{alt, cop, ali, cla\}$).

iteração:

1. determinação da distribuição de *cop*:
   $P(cop = denso) = \alpha_1.P(cop = denso|alt < 10^4 \text{ ft.})$
   $P(cop = esparso) = \alpha_1.P(cop = esparso|alt \geq 10^4 \text{ ft.})$

2. sorteio de um novo estado corrente para *cop* em função das probabilidades calculadas em **1**. Suponha-se que o resultado seria $[cop = denso]$.

3. determinação da distribuição de *cla*:
   $P(cla =presença) = \alpha_2.P(cla =presença|cop = denso, ali = elevado)$
   ...

4. sorteio de um novo estado corrente para *cla* em função das probabilidades calculadas em **3**.

5. executar uma nova iteração.

O algoritmo pára quando o número de iterações inicialmente estipulado tiver sido executado. A distribuição marginal para cada variável não fixada em $\mathcal{E}$ é estimada pela distribuição de frequências das modalidades dessa variável no conjunto de iterações realizadas.

Para as distribuições de probabilidade condicionadas apresentadas na tabela 5.1 (apresentada abaixo) e para 100 iterações as distribuições marginais estimadas foram *P(cla=presença)=0.735, P(cla=ausência)=0.265, P(cop=denso)=0.65* e *P(cop=esparso)=0.35.*

[6] Para aligeirar a notação as variáveis serão identificadas pelas primeiras três letras.

	cop=esparso			cop=denso		
	ali= baixo	ali= médio	ali= elevado	ali= baixo	ali= médio	ali= elevado
cla=ausência	0.67	0.50	0.43	0.67	0.82	0.20
cla=presença	0.33	0.50	0.57	0.33	0.18	0.80

alt$< 10^4$	0.62
alt$\geq 10^4$	0.38

	alt$< 10^4$	alt$\geq 10^4$
cop=esparso	0.58	0.24
cop=denso	0.42	0.76

	cop=esparso	cop=denso
ali=baixo	0.08	0.07
ali=médio	0.68	0.32
ali=elevado	0.24	0.61

Tabela 5.1 - Modelo de incerteza associado ao modelo gráfico da figura 5.1. Na vertical encontram-se as diversas modalidades cujas probabilidades constam das tabelas. Na horizontal encontram-se os estados possíveis para os vértices *pais*.

Podem ser feitos alguns comentários a estes resultados. 1) O modelo de incerteza foi determinado à custa da amostra (da forma descrita em 5.1.2.4) e tem alguns valores que não parecem fazerem muito sentido como, por exemplo, os valores da primeira tabela, para $cop = denso$ e para $alim = baixo$ e $alim = médio$. Essa falha poderia ser corrigida por intervenção do perito. 2) Observando os valores das probalidades e o resultado da propagação da evidência pelo grafo verifica-se que todos os vértices contribuem para todas as distribuições marginais, ou seja, o modelo de representação comporta-se de uma forma global apesar dos cálculos serem efectuados localmente.

O modelo gráfico, a evidência e o resultado da sua propagação podem ser

representados da forma sintética que consta na figura 5.3.

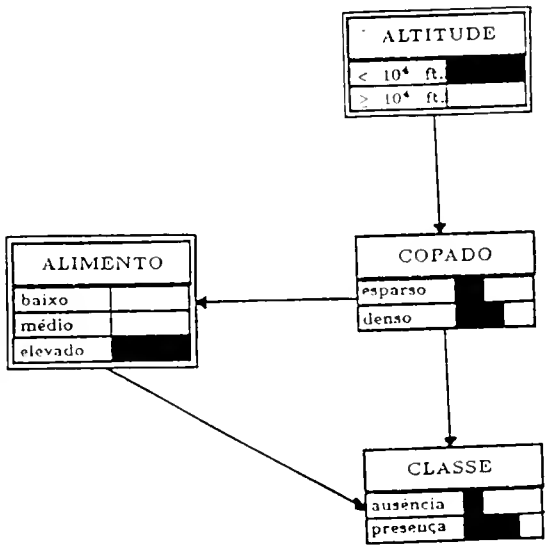


Figura 5.3 - Resultado da propagação de uma evidência pelo grafo

A representação acima é redutora em relação a toda a informação contida no grafo bayesiano mas tem a vantagem de ser muito transparente. O modelo gráfico está representado explicitamente e o modelo de incerteza é representado através do seu efeito global sobre a estrutura. A evidência fixada é facilmente identificável pois corresponde aos vértices em que a distribuição marginal só tem um valor não nulo. O facto dos valores das probabilidades marginais não serem expressos numéricamente não é grave pois os valores, obtidos pelo método 1, são aproximativos. A representação poderia, sem perca de clareza, explicitar esses valores.

6 . Descrição do método de classificação

O processo de aplicação do método de classificação proposto neste trabalho é composto por duas fases. A primeira é a fase de incorporação do conhecimento do perito sobre o domínio do problema no modelo de relações entre as variáveis (o grafo bayesiano). A segunda fase é a da construção da regra de classificação para classificar os objectos de classe desconhecida.

Essas duas fases não serão consideradas isoladamente mas sim como, respectivamente, a fase inicial e a fase final de um mesmo processo, o qual resulta da aplicação do método de classificação a uma instância do problema de classificação. O método é iterativo, pois contém um procedimento que é repetido até que sejam atingidos os objectivos do problema. Esse procedimento consiste na construção de uma árvore de classificação e finaliza-se, portanto, com a descoberta de uma solução para o problema. As primeiras iterações correspondem à primeira fase atrás referida. As últimas iterações correspondem à segunda fase.

No início do processo, o conhecimento pericial não está ainda estruturado, não estando disponíveis as heurísticas (descritas no capítulo 5) baseadas nesse conhecimento. Da forma referida em 5.1.2.4, é construída uma configuração inicial para o modelo de dependências entre as variáveis (grafo bayesiano). Durante a primeira fase o conhecimento pericial é captado (o modelo de dependências entre variáveis vai sendo sucessivamente alterado) e as heurísticas baseadas nesse conhecimento passam a ter um peso crescente na estratégia de construção das árvores de classificação. Na segunda fase (fase em que o modelo de dependências entre variáveis está estabilizado) o conjunto de heurísticas é utilizado para a procura da melhor solução do problema de classificação.

Como se verifica, a árvore de classificação, no método de classificação, não é apenas um suporte da estratégia de procura de soluções mas também é um suporte do processo de incorporação do conhecimento pericial no modelo que representa esse conhecimento (o grafo bayesiano).

Em 6.1 será caracterizado o procedimento de construção de uma árvore de classificação, sendo explicado como se procede à integração das diversas heurísticas descritas nos capítulos 4 e 5. Em 6.2 será descrita a forma de incorporação do conhecimento pericial no modelo. Um aspecto importante, e que não foi desenvolvido anteriormente, é o da avaliação da satisfação dos objectivos apresentados em 2.3. Esse aspecto será desenvolvido em 6.3. Finalmente, em 6.4, será descrita a

implementação do método de classificação. Não é conveniente descrever as noções apresentadas com o exemplo ilustrativo referido nos capítulos anteriores para evitar uma redundância com o capítulo 7, no qual se faz a aplicação do método a um problema real.

6.1. Construção da árvore de classificação: integração das heurísticas

A árvore de classificação é construida através da execução do algoritmo 3.1 que define as ramificações e as afectações às classes. As heurísticas que foram anteriormente descritas auxiliam as tomadas de decisão nesse algoritmo. Na tabela 6.1, apresentada em seguida, relacionam-se as heurísticas com as decisões a tomar.

decisões no nó (algoritmo 3.1)	heurísticas	
	baseadas na amostra	baseadas no grafo bayesiano
(i)-ramificar?	.função informação .proporção de objectos	.independência condicional .probabilidade da classe
(ii)-afectar a uma classe	.proporção de objectos	.probabilidade da classe
(iii)-ramificar por uma variável	.função informação	.independência condicional

Tabela 6.1 - Heurísticas disponíveis para as tomadas de decisão

Antes de caracterizar o papel das várias heurísticas envolvidas na mesma decisão deve-se considerar que a amostra não se altera ao longo do processo de classificação. A informação que dela se pode retirar apenas depende do desenvolvimento da árvore e das soluções do problema exploradas. O mesmo não acontece em relação à estrutura de representação do conhecimento pericial porque, na primeira fase do processo, as heurísticas baseiam-se numa configuração do grafo ainda não consistente com o conhecimento do perito (em 6.2 será discutida a forma de ultrapassar essa inconsistência).

O perito pode, portanto, atender às heurísticas baseadas no grafo bayesiano de duas formas. Se as regras heurísticas são consideradas válidas pelo perito, isto é, se essas regras forem consistentes com o conhecimento do perito sobre o domínio do problema, então são utilizadas para as tomadas de decisão. Se não são consideradas válidas então será necessário alterar o grafo bayesiano, o que originará novas regras heurísticas.

No caso das heurísticas serem aceites como válidas (situação que caracteriza a segunda fase do processo de classificação), a forma de utilização da informação apresentada na tabela 6.1 é a seguinte:

decisão (i) - A ramificação não se realiza se se verificar alguma das seguintes condições: 1) todas as variáveis disponíveis são condicionalmente independentes da classe; 2) não existem variáveis disponíveis para a ramificação; 3) a função de informação é nula para todas as variáveis associadas ao nó; 4) a subamostra é de dimensão significativa e existe c_k tal que $P_A(c_k|\mathcal{E})$ é superior a α, $(\frac{1}{m} < \alpha \leq 1.0)$; e 5) existe um c_k tal que $P(c_k|\mathcal{E}, \mathcal{G})$ é superior a α, $(\frac{1}{m} < \alpha \leq 1.0)$.

decisão (ii) - O nó é afectado à classe c_k se se verificar: 1) a subamostra é de dimensão significativa e existe c_k tal que $P_A(c_k|\mathcal{E}) = max_i P_A(c_i|\mathcal{E})$; ou 2) existe c_k tal que $P(c_k|\mathcal{E}, \mathcal{G}) = max_i P(c_i|\mathcal{E}\,\mathcal{G})$.

decisão (iii) - A variável responsável pela ramificação do nó é a que, de entre as variáveis disponíveis no nó, verificar simultaneamente as seguintes condições: 1) não é condicionalmente independente da classe; e 2) é a variável que torna mínima a função de informação $E(V_k)$.

Como se verifica, existe uma certa imprecisão na definição das condições em que são aplicadas as heurísticas que avaliam as probabilidades de ocorrência das classes em $\mathcal{E}$. Nenhum valor é definido para limiar de α ($\frac{1}{m}$ é simplesmente o valor mínimo que α pode tomar). Esta opção deve-se ao facto do método ser interactivo e do perito poder decidir, em cada situação em que utilize uma dessas heurísticas, o valor de α. Desta forma permite-se que o perito tenha uma maior possibilidade de escolha das soluções a explorar.

Como foi referido anteriormente o máximo erro esperado que o perito poderá cometer numa decisão como a anterior é igual ao produto de $(1 - \alpha)$ pela probabilidade do nó. No caso da probabilidade do nó ser baixa, então fará sentido diminuir o α e, em caso contrário, utilizar um α próximo de 1. Esta estratégia conduz a uma menor complexidade da árvore (em relação ao caso em que α é sempre próximo de 1.0) pois provoca uma mais rápida paragem da ramificação nos ramos em que existe apenas um pequeno número de elementos da amostra (cujos nós têm, consequentemente, uma probabilidade estimada baixa).

Em determinados casos pode ocorrer uma situação de conflito, isto é, podem existir, pelo menos, duas heurísticas aplicáveis para a tomada de decisão e que conduzem a decisões diferentes. A escolha da heurística é feita segundo a ordem

definida atrás. Designe-se heurística rejeitada a uma heurística aplicável mas não utilizada. Por retrocesso, o processo de procura pode conduzir à exploração de uma maior variedade de soluções. Para tal, basta que sejam identificados os nós da árvore nos quais alguma das decisões é realizada numa situação de conflito entre duas ou mais heurísticas e que o processo de procura retorne a esses nós de forma a poderem ser aplicadas as heurísticas rejeitadas.

Importa, ainda, chamar a atenção para os seguintes aspectos que acentuam a importância da integração do grafo bayesiano no método de classificação.

O grafo bayesiano, ao servir de suporte à formalização do conhecimento do perito, capta esse conhecimento e torna-o disponível a ser utilizado em novos raciocínios. Uma situação particular e interessante é da tomada de uma decisão de afectação num nó associado a uma subamostra vazia ou num nó que não seja mais ramificado mas, para o qual, as heurísticas disponíveis não permitam tomar claramente uma decisão de afectação a uma classe. Nesta situação justifica-se uma alteração da estratégia de procura primeiro em profundidade. Esses nós são analisados apenas no final do procedimento de construção da árvore, o que permite que, no decorrer do procedimento, mais conhecimento pericial seja incorporado no modelo.

No método proposto a estratégia de análise dos nós segue a regra acima descrita.

Quando o perito toma qualquer decisão as decisões anteriores são obrigatoriamente consideradas pois estão incorporadas no grafo bayesiano. Esta característica do método de classificação proposto permite ultrapassar a simplificação, estabelecida na árvore de classificação, segundo a qual as decisões em nós não pertencentes ao mesmo *ramo*, e por isso disjuntos, são tomadas isoladamente.

6.2. Captação e estruturação do conhecimento pericial

O grafo bayesiano constitui um suporte para a formalização dos raciocínios do perito. O conhecimento do perito exprime-se, como foi referido em 2.2, como hipóteses sobre relações entre as classes, os objectos e as variáveis que os descrevem. A construção de uma determinada árvore de classificação para o problema é uma forma de explorar o espaço das soluções e, consequentemente, colocar o perito face a pequenos problemas (as decisões no algoritmo 3.1) de fácil formulação. O raciocínio desenvolvido pelo perito para resolver esses pequenos problemas é captado no grafo bayesiano.

O perito, possivelmente, não tem resposta segura para algumas das questões

que lhe são postas. Para ultrapassar essa limitação o perito é auxiliado nas tomadas de decisão de duas formas: 1) a informação da amostra é apresentada de um modo que auxilia o desenvolvimento de raciocínios de tipo indutivo sobre a questão em análise; 2) o método permite a revisão das crenças do perito, isto é, permite a captação de hipóteses e a sua posterior rejeição ou substituição, suportando assim a representação de modos de raciocínios naturais.

Veja-se então, concretamente, como são colocadas ao perito as questões. Para tal, considere-se uma vez mais o algoritmo 3.1. As decisões que o perito deve tomar, ao construir a árvore de classificação, são, em cada nó criado:

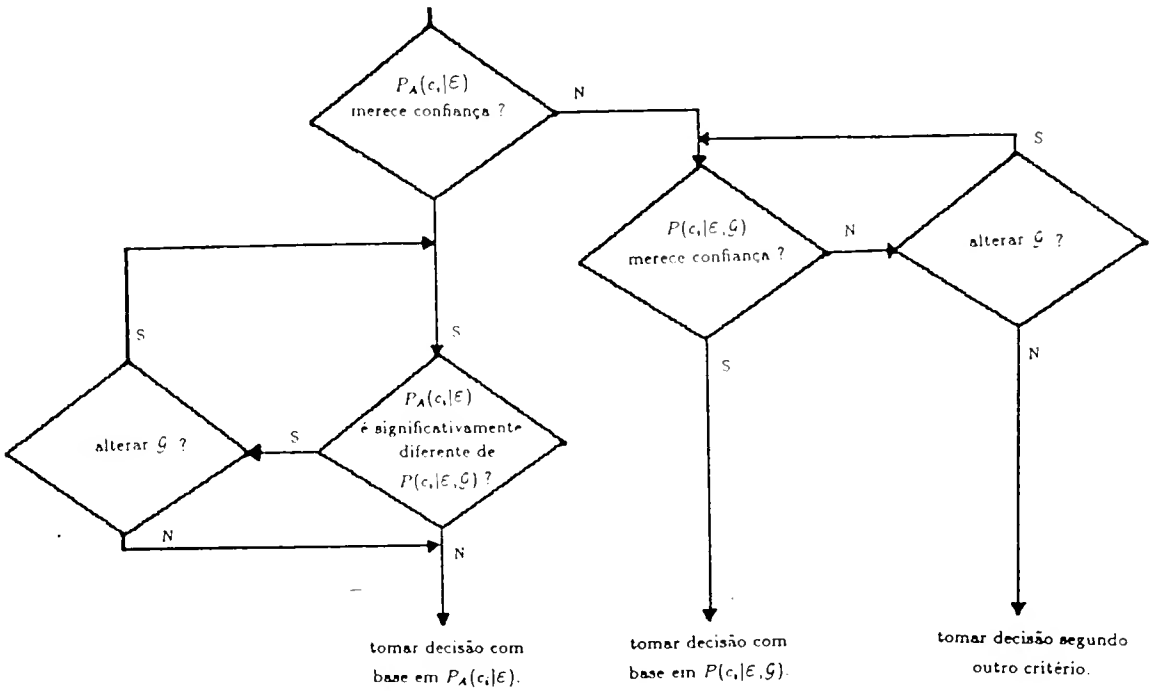
1) Paragem da ramificação e afectação a uma classe - decisões (i) e (ii).

   Esta decisão significa, por um lado, que o perito considera que o estado das variáveis fixadas em $\mathcal{E}$ (a evidência associada ao nó corrente), através da sequência de efeitos que desencadeam sobre outras variáveis e, estas, por seu turno, sobre o vértice *classe*, justifica a afectação a determinada classe. A propagação desses efeitos é formalmente caracterizada no grafo bayesiano. Cada uma das inferências individuais que constituem o raciocínio desenvolvido pelo perito podem ser representadas pelas arestas do modelo gráfico.

   A decisão de afectar um nó, ao qual está associado a evidência $\mathcal{E}$, à classe c_i fundamenta-se, parcialmente, no valor de $P(c_i|\mathcal{E})$. A afectação justifica-se se $P(c_i|\mathcal{E})$ não fôr significativamente diferente de 1 ou se, caso não haja possibilidade de aumentar essa probabilidade por ramificação da árvore, fôr superior a $P(c_k|\mathcal{E})$, $\forall_{k \neq i}$. Existem (v. 6.1) duas estimativas para esse valor.

   A utilização das duas estimativas de $P(c_i|\mathcal{E})$ pode ser integrada no método de classificação através da execução do algoritmo 6.1 apresentado em seguida. A execução deste algoritmo pode conduzir à actualização do grafo bayesiano e,

portanto, à incorporação de conhecimento pericial.



Algoritmo 6.1 - Alteração do grafo bayesiano e decisões (i) e (ii) no algoritmo 3.1

2) Escolha da variável responsável pela ramificação seguinte - decisões (i) e (iii).

Esta decisão deve respeitar a seguinte proposição que resulta da integração da árvore de classificação com o grafo bayesiano e da proposição 5.1.

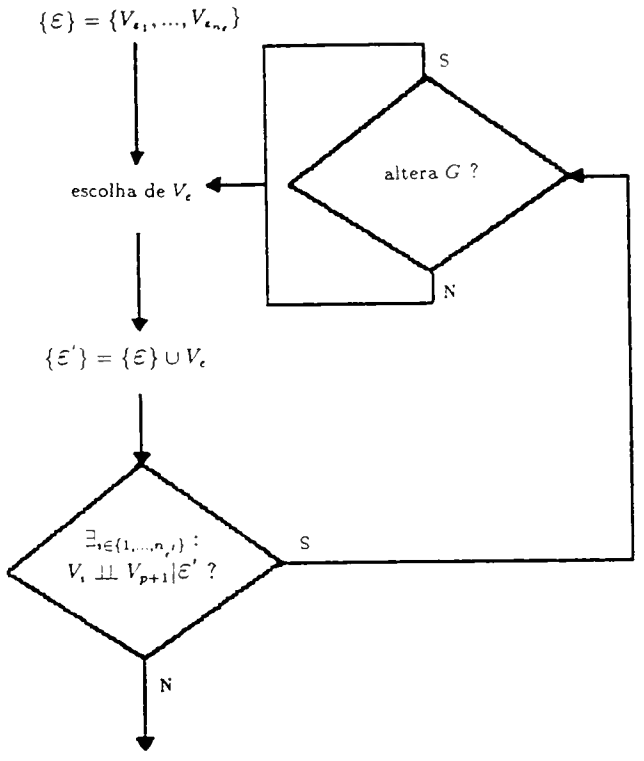
proposição 6.1 Seja $\mathcal{E}$ a evidência associada ao nó corrente da árvore e $\{V_{e_1}, ..., V_{e_{n_e}}\}$ o conjunto de variáveis fixadas em $\mathcal{E}$. Então deve-se verificar a seguinte condição:

$$\nexists_{i \in \{1,...,n_e\}} : V_i \perp\!\!\!\perp V_{p+1} | \{V_{e_1}, ..., V_{e_{n_e}}\},$$

sendo V_{p+1} o vértice *classe* do grafo G.

Este resultado é garantido se fôr executado o algoritmo 6.2, apresentado em seguida, para a escolha da variável responsável pela ramificação do nó

corrente da árvore.



Algoritmo 6.2 - Alteração do grafo bayesiano e decisões (i) e (iii) do algoritmo 3.1

Por um lado, o algoritmo 6.2 pode levar à alteração de G, ou seja, a uma actualização do grafo bayesiano e consequente incorporação de conhecimento pericial na base de conhecimento. Por outro lado, a decisão de ramificar o nó corrente significa que o perito considera que $\mathcal{E}$ não é suficiente por si para justificar uma afectação e que o conhecimento do valor da variável escolhida poderá contribuir a resolver o problema de classificação. A variável escolhida vai aumentar a cadeia de inferências já realizadas na definição do nó.

Foram apresentadas atrás as questões postas ao perito na fase de incorporação do conhecimento pericial no modelo. Caracterizem-se agora, mais pormenorizadamente, as inconsistências que se podem verificar no conjunto do conhecimento envolvido no processo de classificação.

A necessidade de modificar os parâmetros do grafo bayesiano, ou seja, adicionar ou retirar arestas (sem, no entanto, violar a condição de não ciclicidade do grafo), modificar determinadas distribuições de probabilidade, verifica-se quando informação proveniente de

(sa) subamostra do nó corrente da árvore,

(ic) independência condicional entre variáveis do grafo,

(dm) distribuição marginal das variáveis resultante da propagação da evidência correspondente a esse nó,

(cp) conhecimento pericial (não incorporado no modelo)

não for consistente.

As situações em que pode ocorrer inconsistência e as decisões a tomar nessas situações são [1]:

- $(sa \bowtie ic)$ Esta situação corresponde à alteração do grafo bayesiano no algoritmo 6.2.

- $(sa \bowtie dm)$ Esta situação corresponde à alteração do grafo bayesiano no algoritmo 6.1 no caso da inconsistência envolver apenas o vértice *classe*. Este tipo de inconsistência pode resultar de diferenças significativas entre as estimativas das distribuições marginais de outras variáveis. O perito pode decidir alterar o grafo bayesiano por forma a que as diferenças se tornem aceitáveis caso considere que a amostra é representativa. Ao alterar o grafo estará a introduzir o seu conhecimento na base de conhecimento.

- $(sa \bowtie cp)$ Este tipo de inconsistência ocorre quando o perito pretende não escolher como variável responsável pela ramificação da árvore a variável que optimiza a função de informação ou quando o perito não concorda com os critérios de paragem da ramificação descritos em 4.2.

- $(ic \bowtie dm)$ Neste caso a inconsistência, caso se verifique, não é detectável directamente. É corrigida, indirectamente, em outras situações de inconsistência aqui descritas.

- $(ic \bowtie cp)$ Como o modelo gráfico do grafo bayesiano é construído pelo perito não se deveriam verificar inconsistências deste tipo. No entanto, pelo carácter local da construção do modelo gráfico, o perito pode não se aperceber de certas relações indirectas resultantes entre as variáveis. Ao ser confrontado com relações de independência não esperadas o perito pode corrigir essas falhas alterando o modelo gráfico.

[1] Entre parêntesis é indicada a origem da inconsistência. O símbolo $\bowtie$ significa *inconsistente com.*

- ($dm \bowtie cp$) O perito pode não concordar com os valores das probabilidades marginais de uma ou mais variáveis obtidos por propagação da evidência corrente pelo grafo. Caso se verifique inconsistência o perito deve alterar o modelo de incerteza (alterando uma ou mais distribuições de probabilidades condicionadas) e, eventualmente, o modelo gráfico.

O processo de actualização da base de conhecimento descrito acima processa-se sequencialmente segundo a ordem de criação dos nós da árvore de classificação. Esta característica é, por um lado, vantajosa pois define uma ordem de processamento e um enquadramento preciso para o raciocínio do perito mas leva, por outro lado, a uma análise *local* do domínio dos objectos que pode levar o perito a tomar decisões válidas para uma determinada região de Ω' mas não generalizáveis a todo o Ω'. Este aspecto não constitui um problema no que respeita à árvore de classificação (pois as regiões de decisão são disjuntas) mas deve ser tomado em consideração no que respeita o grafo bayesiano. De facto, supõe-se que o modelo gráfico e o modelo de incerteza são válidos em todo o Ω'. Esse problema é ultrapassado por se repetir iterativamente o procedimento de construção da árvore de classificação garantindo-se desta forma que toda a informação existente sobre Ω' contribui para a resolução do problema.

6.3. Avaliação da satisfação dos objectivos do método

Recorde-se que os objectivos (apresentados em 2.3) que se pretende atingir com o método de classificação proposto são os seguintes: 1) construir uma regra de classificação que permita classificar correctamente os objectos de classe desconhecida; 2) tirar partido de toda a informação disponível (e, em particular, do conhecimento pericial) para construir essa regra; e 3) captar a estrutura preditiva para a classificação das variáveis envolvidas e construir uma base de conhecimento sobre o domínio do problema.

Relativamente ao primeiro objectivo, é de esperar que, na segunda fase do processo de classificação, quando o modelo do conhecimento pericial está concluído, o desempenho da regra de classificação seja superior ao do método ID3 pois é construida segundo as regras heurísticas de ID3 quando essas regras merecem confiança e é construida segundo o modelo de representação do conhecimento pericial quando as regras baseadas na amostra não merecem confiança. Uma forma objectiva de controlar do erro cometido na classificação de objectos de classe desconhecida, por

utilização da regra de classificação obtida, consiste em calcular uma taxa de erro aparente (razão entre o número de objectos da amostra que são classificados erradamente e o número total de objectos da amostra). A estimativa resultante, para o erro de classificação, é geralmente optimista. Outra forma baseia-se numa técnica de validação cruzada. Esta técnica consiste em utilizar como amostra de treino da árvore apenas uma parte do conjunto de objectos de classe conhecida. A "qualidade" da regra é determinada pela aplicação da regra aos restantes objectos. Esta técnica dá bons resultados, isto é, dá estimativas do erro próximas do valor da taxa real de má-classificação (cf. Gnanadesikan *et al.*, 1989, Breiman *et al.*, 1984 e Tomassone, 1988).

A procura da melhor, segundo algum dos critérios acima apresentados, regra de classificação é feita por exploração de diversas soluções na árvore de classificação e por cálculo do valor da taxa de erro para cada solução. Em alternativa ao cálculo dessa taxa podem ser utilizadas funções heurísticas (v. 3.2). As heurísticas $P_A(c_i|\mathcal{E})$, $P(c_i|\mathcal{E},\mathcal{G})$ e $min_k\{E(V_k)\}$ são também funções heurísticas pois dependem da taxa de erro aparente que resulta da afectação do nó em análise a uma classe e permitem, portanto, avaliar a "qualidade" dessa afectação.

A satisfação do segundo objectivo relaciona-se com o grau de estruturação do conhecimento pericial sobre o domínio do problema. O grau de estruturação máximo, no método proposto, corresponde à configuração do grafo bayesiano que exprime da forma mais correcta o conhecimento do perito. Esse grau considera-se atingido quando a primeira fase do processo de classificação está concluida, ou seja, quando o perito considera que o grafo bayesiano não deve ser alterado.

A satisfação só poderia ser total se o modelo do conhecimento pericial e as heurísticas definidas sobre esse modelo representassem todo o conhecimento do perito relevante para resolver o problema. O resultado da aplicação do método proposto poderá aproximar-se mais ou menos dessa situação em função das características do problema.

Em relação ao terceiro objectivo pode dizer-se que a satisfação deste depende, por um lado, da satisfação do segundo objectivo (porque, no método, tirar partido do conhecimento pericial implica a sua prévia incorporação numa estrutura que, no final do processo, se torna uma componente da base de conhecimento) e, por outro lado, da interpretabilidade da árvore de classificação (a outra componente dessa base de conhecimento). A árvore de classificação constitui um suporte simples para a regra de classificação e o grafo bayesiano permite representar uma maior diversidade de

conhecimento, em particular, as relações de dependência entre as variáveis e entre estas e as classes.

A interpretabilidade está relacionada com a complexidade da árvore. Desta relação considerem-se duas consequências importantes. Por um lado, o método proposto produz árvores menos complexas que ID3 porque tem critérios mais flexíveis para a paragem de ramificação e, portanto, origina regras mais interpretáveis. Por outro lado, dessa relação conclui-se que o primeiro e o terceiro objectivos podem ser conflituosos pois uma maior complexidade da árvore pode, eventualmente, conduzir a uma menor taxa de erro de classificação e tornar a regra de classificação menos interpretável.

6.4. Descrição do programa

O programa de classificação, que será designado por PCIC (programa de classificação e integração de conhecimento).

6.4.1. Arquitectura do programa

O programa é constituido por cinco módulos principais: **freq_rede**, **iteracao**, **arvore**, **verif_indep** e **propag** (os três últimos estão incluidos em **iteracao**). **freq_rede** é executado no início do processo e tem como finalidade produzir uma versão inicial do modelo de incerteza associado ao grafo (da forma descrita em 5.1.2.4). **iteracao** é o módulo que inclui praticamente todo o programa e que é responsável pela iteratividade. O critério de paragem de **iteracao** é a estabilização da base de conhecimento (grafo bayesiano e árvore de classificação). O módulo **arvore** é o mais importante e é o responsável pela construção da árvore de classificação. É um módulo recursivo pois é executado da mesma forma para cada novo nó criado. **verif_indep** determina o estado de dependência ou independência das variáveis não fixadas relativamente à classe, dada a evidência corrente. O módulo **propag** é responsável pela propagação, segundo o método estocástico proposto por Pearl (1987), da evidência corrente pelo grafo e pelo cálculo das distribuições de probabilidade marginais nos nós do grafo.

Existe ainda um módulo que não faz parte de PCIC mas que é utilizado, no exemplo apresentado, para calcular a taxa de erro real da regra de classificação. Trata-se do módulo **classifica** que calcula, a partir da totalidade dos objectos da população considerada, as diversas proporções de objectos bem e mal classificados

pela regra de classificação.

A entrada de PCIC é constituida por: 1) amostra; 2) modelo gráfico inicial do grafo bayesiano; 3) definição das partições consideradas para os domínios das variáveis; 4) estado inicial para os nós do grafo. O saída consiste na árvore de classificação e no grafo bayesiano (componente gráfica e componente probabilística).

Na figura 6.1 é apresentado um fluxograma onde constam os módulos principais do programa. São também indicados (com o símbolo #) os passos do processo em que o perito intervém na tomada de decisão e em que tem oportunidade de alterar os parâmetros do grafo bayesiano. Algumas características adicionais são apresentadas na discussão da complexidade computacional. O programa PCIC é apresentado em anexo (anexo B).

6.4.2. Complexidade computacional

Como foi referido PCIC é constituido por um conjunto de módulos. Tente-se analisar a complexidade de cada um deles e o número de vezes que cada módulo é executado.

O módulo **freq_rede** é executado uma única vez. Como foi referido em 5.1.2.4, o número de valores a estimar em **freq_rede** é $\sum_{i=1}^{p}(n_{V_i} \cdot \prod_{j=1}^{npa_i} n_{pa_j(V_i)})$. Embora esse número cresca exponencialmente com npa_i (o número da pais do vértice V_i do grafo), a construção e forma de execução do método garantem que não se põe o problema de um aumento exponencial do tempo de cálculo. Se o número atrás referido fôr elevado (porque existe, por exemplo, alguma variável dependente, segundo o modelo gráfico, de um grande número de outras variáveis), o perito deixa de poder controlar o modelo de incerteza e o método proposto deixa de ser interessante e deve ser abandonado.

O módulo **iteracao** é executado um número finito (e pequeno) de vezes e, portanto, não aumenta a ordem da complexidade computacional do método.

Em cada iteração o módulo **arvore** é executado para cada nó analisado da árvore. O número de nós não é superior ao produto do número de objectos da amostra (L, número máximo de nós terminais) pelo número de variáveis (p, número máximo de níveis da árvore) .

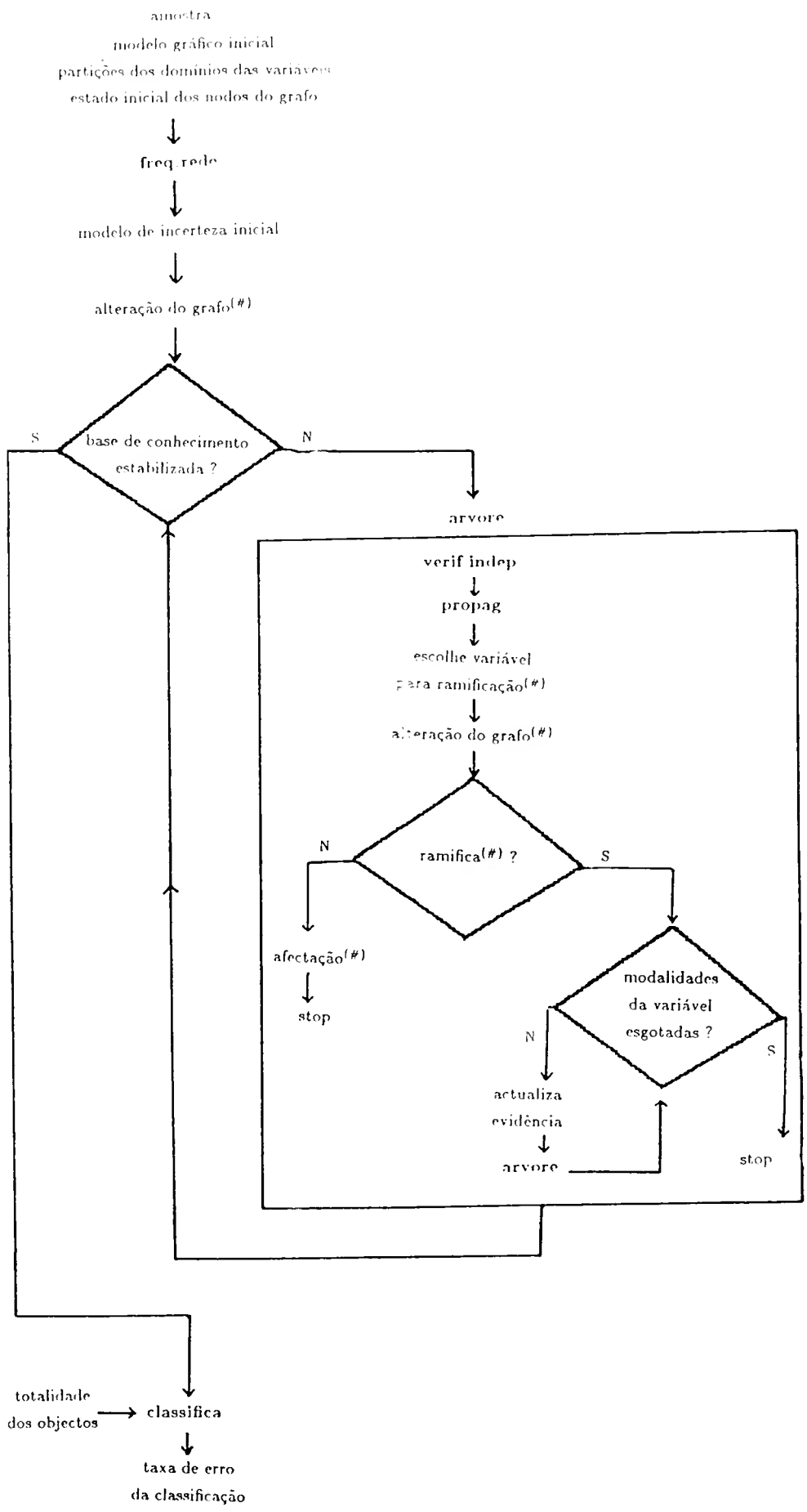


Figura 6.1 - Fluxograma do programa PCIC

Cada vez que um nó é analisado são executados os seguintes algoritmos. 1) Módulo **verif_indep** - que, dado um vértice (*classe*), e um conjunto S de vértices (as variáveis que definem a evidência), determina o conjunto de vértices d-separados por S de *classe*. A implementação realizada consistiu num algoritmo baseado numa procura exaustiva dos caminhos, para o qual a complexidade não é polinomial no número de arestas. Dada a natureza desse problema é possível que exista um algoritmo polinomial para o resolver. 2) Módulo **propag** - que, em cada iteração (o número de iterações para a estimação da distribuição marginal de probabilidades é constante), deve, para cada variável, calcular o produto que consta na proposição 5.3. A complexidade de **propag** é $O(p+|E|)$ segundo (Pearl, 1987). 3) Determinação do valor da função de informação para, no máximo, as p variáveis. A determinação de $E(V_k)$ exige, para um número inferior a L objectos, a contagem da proporção de objectos em cada modalidade de V_k (o número de modalidades é n_k) e em cada classe (v. 4.1). A complexidade do algoritmo para determinação da função informação é, então, $O(p.L.nmod.m)$, sendo $nmod = max_k(n_k)$.

No caso do perito pretender alterar parâmetros do grafo bayesiano, os módulos **propag** e **verif_indep** podem ser executados mais do que uma vez durante a execução do módulo **arvore**. Este facto não aumenta a ordem de complexidade.

A complexidade computacional de PCIC não foi determinada mas, se existir um algoritmo polinimial para **verif_indep**, **iteração** terá complexidade polinomial. Pelas razões apresentadas para a discussão da complexidade de **freq_rede** e pela interactividade de MCIC é possível que tempo necessário para a execução não seja um factor limitante da utilização de PCIC.

6.4.3. Linguagem de programação

A linguagem utilizada para a implementação foi a linguagem PROLOG, dialecto C-PROLOG e versão 1.2. O processamento foi realizado num computador VAX 4200. O dialecto e a versão utilizadas são antigos e já ultrapassados. Este aspecto não constituiu uma limitação para este trabalho porque a aplicação exemplificativa realizada (que será apresentada no próximo capítulo) envolveu um problema real de pequena dimensão.

O PROLOG é uma linguagem de programação em lógica. É um método de resolução de problemas e simultaneamente, por ser uma linguagem de comunicação com os computadores, um instrumento de programação (Coelho *et al.*, 1992). Relativamente ao problema de classificação, o processo de procura do sistema PROLOG segue a mesma estratégia que é adoptada para o método de classificação (procura primeiro

em profundidade), o que torna desnecessária a implementação de um mecanismo de contrôle para a exploração dos subproblemas considerados.

O PROLOG tem um estilo declarativo, que permite uma grande flexibilidade no desenvolvimento de um programa e é adequado à representação e processamento do conhecimento envolvido no processo de classificação.

A eficiência do programa implementado é baixa no que respeita o módulo **propag**. Pelo facto do cálculo, nesse módulo, ser essencialmente numérico (cálculo da propagação de probabilidades pelo grafo bayesiano), **propag** deveria ser implementado numa linguagem mais adequada, nesse aspecto, que o PROLOG.

Ainda sobre o programa PCIC, o número de predicados definidos é 101 e o programa ocupa 48 Kb de espaço de memória. O programa tem cerca de 20 páginas de código. Este último valor deve-se principalmente ao facto de PCIC ser interactivo e necessitar de uma interface pesada.

7 . Aplicação do método a um problema real

O problema considerado neste capítulo é um problema de pequena dimensão e a aplicação realizada visa, principalmente, ilustrar o processo de classificação, ou seja, o processo de incorporação do conhecimento pericial no modelo (grafo bayesiano) e a construção da regra de classificação.

Para além da descrição do processo, será comentado o desempenho do método de classificação proposto (que será designado por MCIC, por analogia com a designação dada ao programa) atendendo aos objectivos expostos em 2.3. Por forma a tornar mais completo esse comentário, os resultados de MCIC serão comparados com os resultados de dois métodos convencionais referidos em 2.4: um método paramétrico (regressão logística) e um método baseado numa árvore (ID3).

7.1. Descrição do problema

Os dados que caracterizam a instância do problema de classificação aqui considerado foram reunidos para um estudo apresentado numa tese de doutoramento (Pereira, 1989). Uma das finalidades desse estudo era a construção de um modelo para a previsão da localização dos nichos ecológicos da espécie *Tamiasciurus hudsonicus grahamensis* em Mt. Graham, uma região montanhosa no Arizona, E.U.A.. A motivação ecológica para a construção desse modelo era a de prever o efeito negativo que a população dos esquilos poderia sofrer com a instalação de um importante observatório astronómico numa zona da região considerada.

Esse problema era considerado importante porque, por um lado, a espécie considerada é autóctone da região e tem um grande valor patrimonial e turístico e, por outro, a região reúne condições ímpares para a instalação do observatório e esse observatório estava integrado num importante projecto de observação astronómica.

Os objectos considerados são parcelas de terreno (com uma área unitária de 0.5 ha., isto é, 5000 m^2). As variáveis consideradas em (Pereira, 1989) para a caracterização dessas parcelas são descritas na tabela 7.1 [1]. Nessa tabela consta, para cada variável, a sua designação, o seu significado, o seu tipo (variável contínua ou discreta), o conjunto de valores que pode tomar e as unidades em que esses valores são expressos. Algumas unidades são do sistema inglês (as unidades de distância ft.

[1] Essas variáveis serão, quando necessário para simplificar a notação, substituídas pelas quatro primeiras letras (*alti* em vez de *altitude*, por exemplo).

e in.).

variável	descrição	tipo	domínio	unidades
alti	altitude	contínua	[6787,10680]	ft.
decl	declive	contínua	[0,180]	percentagem
dist	distância a clareiras	contínua	[0,31]	nºparcelas
alim	nível alimentar	discreta	{nulo, baixo, médio, alto}	nominal
expo	exposição ao sol	contínua	[0,360]	graus
copa	densidade de copado	discreta	{[0,10],[10,40],[40,70],[70,100]}	percentagem
diam	diâmetro médio dos troncos	discreta	$\{0,]0,5], [5,9], \geq 9\}$	in.

Tabela 7.1 - Descrição das variáveis

Os objectos podem pertencer a uma de duas classes: a classe *ausência* que corresponde à ausência de nichos de um esquilo na parcela e a classe *presença* que corresponde à presença de nichos do esquilo na parcela.

Os valores das variáveis foram obtidos para todas as parcelas através de cartas e bases de dados dos serviços florestais. A presença ou ausência de nichos foi determinada directamente no terreno. Por ser conhecida a verdadeira classe a que pertence cada objecto é possível validar a regra de classificação, calculando a taxa de erro real.

O problema que será considerado neste capítulo consiste em classificar 471 desses objectos, 259 dos quais pertencentes à classe *ausência* e 212 pertencentes à classe *presença*. Foram utilizadas, como entrada, amostras de treino escolhidas aleatoriamente com dimensões de 25 e 50 objectos (aproximadamente 5% e 10% do número de elementos da população)

Em seguida, são apresentadas algumas informações que dizem respeito aos hábitos ecológicos dos esquilos da zona considerada e às variáveis referidas. Essas informações exprimem parte do conhecimento pericial sobre o domínio do problema e serão, como se verá em 7.2.2, incorporadas no modelo de dependências entre as variáveis.

Os esquilos preferem locais húmidos e sombrios, precisam de fontes da alimento (próximas) e de condições de elevadas dimensões (de preferência, diâmetro superior a 12"), elevadas densidades de copado, e grande agrupamento de árvores. A região considerada está no extremo sul da zona ecológica dos esquilos e das espécies arbóreas das quais eles dependem, o que leva os esquilos a serem bastante selectivos na escolha dos seus nichos. As variáveis altitude e distância a clareiras podem estar

negativamente correlacionadas porque as estradas existentes (consideradas como casos particulares de clareiras) foram construídas em locais altos. O efeito (directo) da distância sobre a actividade dos esquilos não é bem conhecido (à escala considerada pode até não ter efeito nenhum). As espécias arbóreas que fornecem boas condições aos esquilos desenvolvem-se melhor nas zonas altas.

O conjunto de dados do problema disponíveis não são dados brutos mas sim dados que já tinham sofrido algumas transformações. Este facto merece algumas considerações. Por um lado, tinha havido a preocupação de transformar e codificar a informação disponível por forma a poder utilizar técnicas estatísticas convencionais. Os dados disponíveis para a exemplificação do método MCIC resultam dessas modificações. Por outro, a informação original estava sobre forma de cartas e bases de dados em que haviam sido introduzidas reduções e codificações dos dados.

Teria sido mais interessante, para este trabalho, dispôr de dados num estado mais *bruto*. Por exemplo, em vez de ter um índice de nível alimentar seria útil saber que espécias arbóreas (algumas das quais são preferidas pela espécie de esquilos) existiam nas parcelas, o que permitiria que o perito, ao fazer as simplificações necessárias, mantivesse discriminada a informação pertinente para definir as suas hipóteses de trabalho.

7.2. Classificação

Nesta secção serão apesentadas as aplicações dos três métodos referidos no início do capítulo ao problema colocado. Em 7.2.1 serão considerados os métodos de classificação totalmente induzida pela amostra: regressão logística (método estatístico clássico) e ID3. Em 7.2.2 será descrita a aplicação de MCIC através da apresentação de alguns passos do processo de classificação. A descrição da totalidade do processo para a amostra de 25 objectos pode ser encontrada em anexo (anexo A).

7.2.1. Métodos induzidos de classificação

Foram aplicados dois métodos de classificação automática, induzida pela amostra. O primeiro exige a codificação das variáveis discretas (categoriais) através de variáveis binárias. Este primeiro método (regressão logística) foi implementado na linguagem GENSTAT5.

O segundo método, que é baseado numa árvore não binária, implica a discretização das variáveis contínuas envolvidas. ID3 foi implementado em C-PROLOG.

7.2.1.1. Método paramétrico

O método comparativo é a regressão logística com definição de um ponto de corte entre as duas classes. São estimados por regressão múltipla os parâmetros de um modelo linear generalizado cuja função associada é a função logística, supondo uma distribuição binomial dos resíduos (cf. Tomassone *et al.*, 1988). A regra de classificação obtida pertence ao grupo de regras baseadas em funções discriminantes e é adequada a situações em que a função de densidade conjunta de probabilidade do conjunto de variáveis em cada população não é normal multivariada. Escolhem-se as variáveis a incluir no modelo através de um método iterativo (*forward stepwise*).

A regressão logística foi utilizado em (Pereira, 1989) para o mesmo problema, mas para uma amostra de maior dimensão (aproximadamente 380 objectos), e levou à obtenção de uma regra de classificação que foi considerada como sendo um bom modelo de predição.

Vários modelos podem ser obtidos em função do critério de paragem do método iterativo de inclusão de novas variáveis. O critério escolhido, neste caso, impede a inclusão de variáveis pouco significativas e origina, portanto, modelos parcimoniosos. Por se tratar de um caso académico e por se poderem calcular as taxas de erro reais foram experimentados vários modelos e concluiu-se que o modelo parcimonioso tinha uma taxa de erro real inferior à do modelo que classificava melhor a amostra.

Para uma amostra de 25 objectos a regra de classificação obtida é

$$\frac{1}{1+e^{-(-91.3+0.00949.altitude-1.86.distância)}} > 0.75 \Rightarrow classe = presença$$

e é, portanto, equivalente a

$$0.00949.altitude - 1.86.distância > 92.398 \Rightarrow classe = presença.$$

A superficie de decisão é um hiperplano no espaço das variáveis e define apenas duas regiões de decisão. Segundo a definição proposta em 3.4, a complexidade da regra é, portanto, 2. A regra de classificação classificou na classe correcta 72.6% dos 471 objectos para a amostra de dimensão 50 e 71.1% para a amostra de dimensão 25. A informação qualitativa que se pode retirar desta expressão para caracterizar o domínio dos objectos é da forma: *se a altitude aumentar ou se a distância diminuir então a probabilidade de existirem esquilos aumentará.*

7.2.1.2. Método baseado numa árvore

O método utilizado foi o ID3 (método referido em 2.4 e 3.1).

Um dos aspectos a considerar é a discretização das variáveis contínuas. É necessário definir, para essas variáveis, partições finitas dos seus domínios. Essas partições podem ser escolhidas por observação, para cada variável, da distribuição dos objectos das classes pelo domínio da variável, por forma a identificar as regiões, em cada domínio, nas quais uma das classes predomina.

Para automatizar o processo de determinação das modalidades das variáveis foi construído um programa para determinar as curvas de frequências acumuladas de objectos para cada uma das classes (e para cada variável) e para identificar os pontos do domínio em que a diferença entre as curvas de frequência atingia um máximo relativo. As duas abordagens conduziram a resultados semelhantes. Houve a preocupação de escolher o menor número possível de modalidades para cada variável por forma a limitar a complexidade da árvore . As modalidades escolhidas para cada variável são apresentadas na tabela 7.2. Nessa tabela define-se, quando necessário, o significado das modalidades consideradas como função da definição apresentada na tabela 7.1.

variável	modalidades	significado
altitude	$< 10^4$ ft.	-
	$\geq 10^4$ ft.	-
declive	$< 20\%$	-
	20 a 40 %	-
	$\geq 40\%$	-
distância	$< 5 parcelas$	-
	$\geq 5 parcelas$	-
alimento	baixo	nulo e baixo
	médio	-
	elevado	-
exposição	norte	0 a 90 graus
	este	90 a 180 graus
	sul	180 a 270 graus
	oeste	270 a 360 graus
copado	esparso	0 a 40%
	denso	40 a 100%
diâmetro	$< 9''$	-
	$\geq 9''$	-

Tabela 7.2 - Modalidades consideradas para as variáveis

A árvore induzida pela amostra de 25 objectos é constituida pelo seguinte conjunto de regras:

$[alti \geq 10^4 \text{ft.}] \wedge [dist \geq 5 parc.] \Rightarrow clas = ausência;$

$[alti \geq 10^4\text{ft.}] \wedge [dist < 5parc.] \Rightarrow clas = presença;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [decl \geq 40\%] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [20\% \leq decl < 40\%] \wedge [expo = norte] \Rightarrow clas = presença;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [20\% \leq decl < 40\%] \wedge [expo = este] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [20\% \leq decl < 40\%] \wedge [expo = sul] \Rightarrow clas = não\ classificado;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [20\% \leq decl < 40\%] \wedge [expo = oeste] \Rightarrow clas = não\ classificado;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \wedge [decl < 20\%] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = médio] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = baixo] \Rightarrow clas = não\ classificado.$

Este conjunto de regras, quando aplicadas à totalidade dos 471 objectos, classifica na classe correcta 66.9% dos objectos e não classifica 11.0%. Aplicando ID3 à amostra de 50 objectos esses valores passam a ser, respectivamente, 73.0 % e 3.9%.

A complexidade da regra de classificação obtida para a amostra de 25 objectos é de 10 (pois origina dez regiões de decisão). Para a amostra de 50 objectos a árvore é muito mais ramificada e a complexidade atinge 21. Aliás, verificou-se, para o mesmo problema, que a complexidade de ID3 cresceu linearmente para amostras de dimensão compreendida entre 20 e 50 objectos.

Verifica-se que o conhecimento captado sobre o domínio dos objectos é constituido por um conjunto de regras de interpretação relativamente fácil que descrevem com maior pormenor o tipo de relações existentes entre variáveis e classes do que a regra de classificação obtida pelo método de regressão logística. Por exemplo, compreende-se que para altitudes elevadas a presença ou ausência depende da distância às clareiras mas que para altitudes baixas os factores mais importantes são outros. A quarta regra, quando comparada com outras regras aplicáveis a parcelas de baixa altitude, exprime a hipótese de que em parcelas de baixa altitude mas com outras condições ambientais favoráveis (alimento elevado, nomeadamente) a exposição é um factor determinante para a presença de nichos de esquilos. Trata-se de uma hipótese sobre o domínio do problema que é interessante pois, de facto, faz sentido supôr que os esquilos preferem zonas frescas e, portanto, expostas a norte.

7.2.2. Método de classificação baseado em conhecimento

Um processo de classificação para a amostra de 25 objectos irá ser descrito por algumas fases de construção da árvore de classificação e de alteração do grafo bayesiano. O processo é descrito na sua totalidade e em pormenor no anexo A.

Nesse anexo estão identificados, pelos números das figuras que serão apresentadas em seguida, os passos do processo de classificação comentados por essas figuras, por forma a poder ser seguida, com maior pormenor, a interacção entre o perito e o programa.

As partições associadas aos domínios das variáveis são as apresentadas na tabela 7.2. O modelo gráfico inicial é o que será apresentado para ilustrar a primeira fase do processo. O estado inicial dos nodos do grafo escolhido é, para cada um, o estado mais frequente na amostra.

O número de iterações fixado para o algoritmo de propagação de evidências foi determinado por forma a garantir a convergência do algoritmo de propagação. O algoritmo foi executado com um número de iterações de 100, 200 e 300. Para avaliar a convergência foram comparados os resultados para quatro estados iniciais para as variáveis: o estado mais frequente na amostra, o estado menos semelhante com este, e outros dois determinados aleatoriamente. O valor máximo da diferença entre estimativas da mesma probabilidade marginal nos quatro casos foi 0.11 (para 100 iterações), 0.08 (para 200 iterações) e 0.02 (para 300 iterações). Considerou-se que o valor para o último caso era aceitável e, portanto, o número de iterações escolhido foi 300.

O modelo de incerteza inicial é determinado a partir da amostra (como referido em 5.1.2.4), com base no modelo gráfico inicial. De modo a evitar a existência de valores nulos para as probabilidades condicionadas (que seriam frequentes, dada a dimensão reduzida da amostra) estabeleceu-se um valor mínimo que não excedeu, em nenhum caso, 0.02.

Definida a entrada do programa, e determinado um estado inicial para o modelo de incerteza (que o perito poderia ter alterado à partida, se desejasse, não sendo nesse caso auxiliado pelo método) o processo de classificação é iniciado.

Para descrever, em cada passo do processo, o grafo bayesiano, será utilizada a representação da figura 5.3. Por forma a facilitar a comparação entre os valores de $P(c_i|\mathcal{E})$ estimados pela amostra e dados pelo grafo bayesiano, os valores provenientes da amostra serão apresentados de uma forma semelhante, diferindo apenas as representações na consideração (valores dados pelo grafo) ou não consideração (frequências na amostra) de arestas entre as variáveis.

O *nodo raiz* representa a amostra. A evidência associada ao *nodo raiz* é vazia e pode ser representada por [ ]. As frequências das modalidades na amostra são

representadas na figura 7.1.

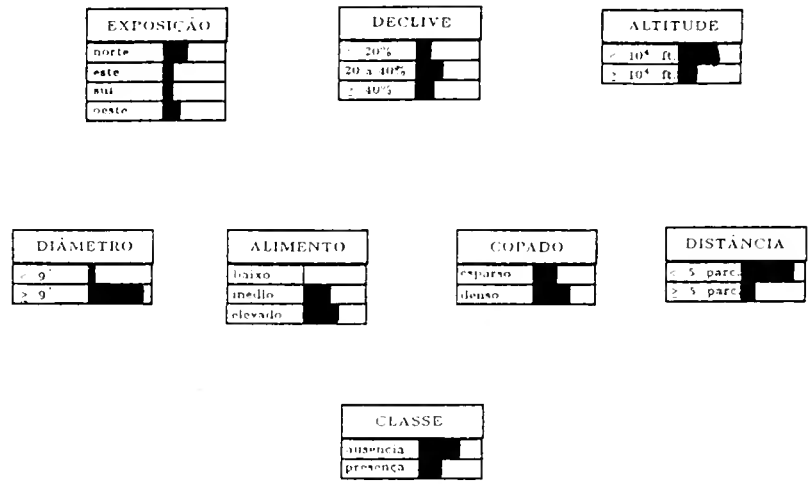


Figura 7.1 - Distribuições de frequências das variáveis na amostra (nó [ ])

Na figura 7.2 apresenta-se o modelo gráfico inicial e as distribuições marginais resultantes da propagação de $\mathcal{E}$=[ ] pelo grafo bayesiano.

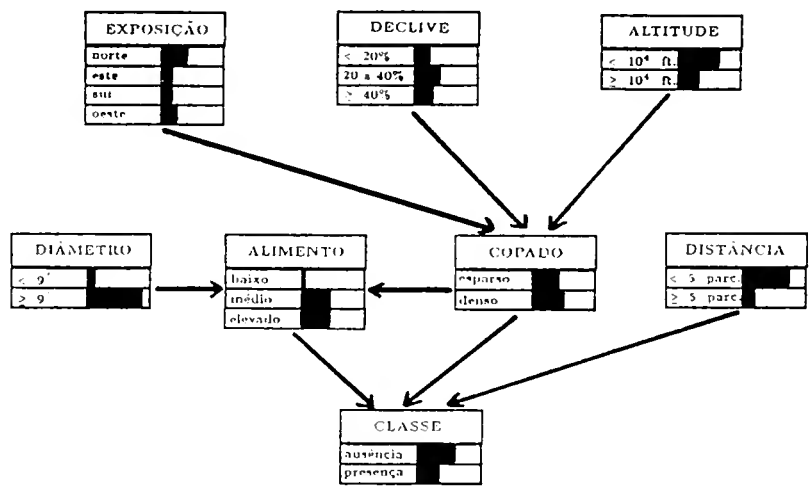


Figura 7.2 - Grafo bayesiano: propagação de $\mathcal{E}$=[ ]

A variável escolhida pelo critério da função de informação é a variável *alti*. O nó corrente passa a ser $\mathcal{E}$=[*alti* $\geq 10^4$]. A esse nodo estão associados 1 objecto da classe *ausência* (11%) e 8 objectos da classe *presença* (89%). A estimativa $P_A(v_i|\mathcal{E})$ e o

valor de $P(v_i|\mathcal{E},\mathcal{G})$, para todas as variáveis, são apresentadas nas figuras 7.3 e 7.4, respectivamente.

Figura 7.3 - Distribuições de frequências das variáveis na subamostra do nó $[alti \geq 10^4]$

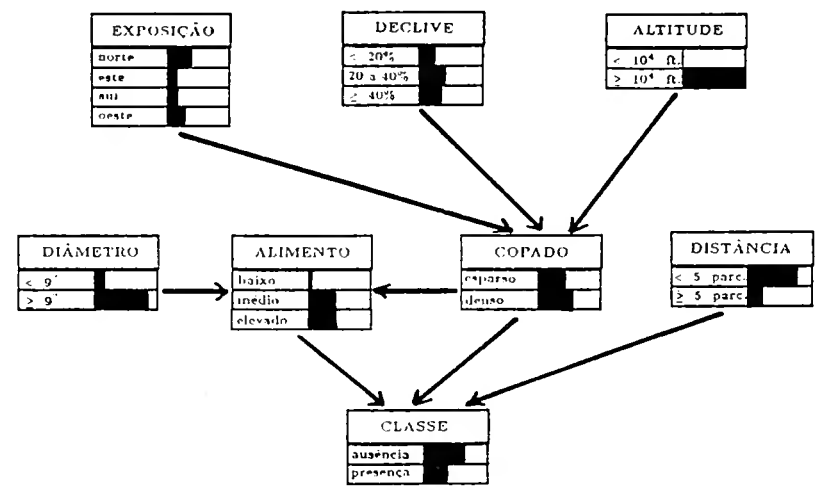


Figura 7.4 - Grafo bayesiano: propagação de $\mathcal{E}$=$[alti \geq 10^4]$

Todas as variáveis continuam a ser independentes do nodo classe.

Como se pode observar no fluxograma de PCIC (figura 6.1) o perito tem a possibilidade de alterar o grafo bayesiano antes da fixação de uma nova evidência. No passo corrente o perito, ao verificar a existência de uma inconsistência do tipo $sa \bowtie dm$ para a variável copado (pois a distribuição das modalidades estimada

na subamostra e dada pelo grafo bayesiano é muito diferente), decide alterar o modelo de incerteza definindo alguns novos valores para a distribuição condicionada $p(copa|alti, decl, expo)$.

As alterações que decide realizar são pequenas pois não encontra disparidades importantes entre o seu conhecimento e os valores das probabilidades estimados para o modelo de incerteza inicial. O algoritmo de propagação é de novo executado para o perito verificar o efeito das alterações introduzidas. O resultado consta da figura 7.5.

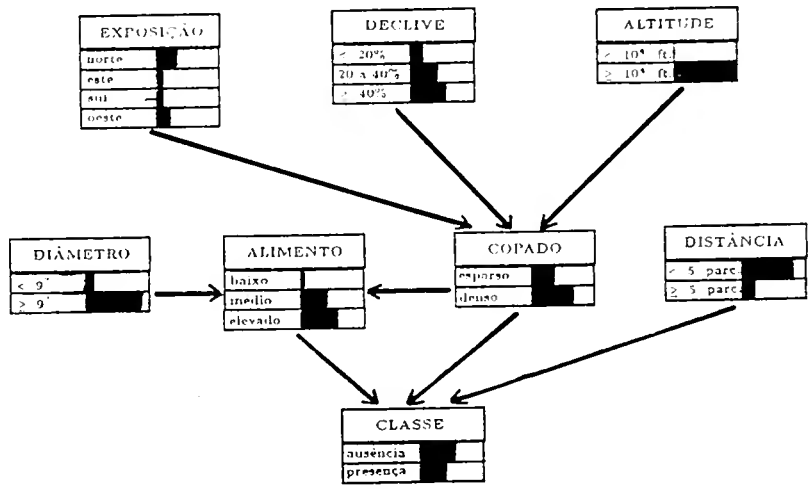


Figura 7.5 - Grafo bayesiano: propagação de $\mathcal{E}=[alti \geq 10^4]$

Observando os resultados o perito observa que a distribuição do vértice *copa* se alterou da forma pretendida mas considera continuar a haver uma inconsistência entre as distribuições marginais dos vértices *clas*, *copa* e *alim* estimadas pela amostra e dadas pelo grafo. A origem dessa inconsistência poderia ser a não consideração de uma relação de dependência directa entre *alim* e outra variável. Sabendo que a produção de frutos pelas árvores locais pode ser influenciada directamente por condições de calor e humidade o perito põe a hipótese de *alim* depender directamente de *alti* (não controlando variáveis como a temperatura, radiação solar incidente ou humidade, o perito considera que a altitude é uma variável que está correlacionada fortemente com esses parâmetros físicos). Decide alterar o grafo adicionando a aresta $alti \rightarrow alim$. Decide verificar também se a distribuição $p(alim|diam, alti, copa)$ obtida é consistente com o seu conhecimento. Na figura 7.6 mostra-se o resultado

da propagação da evidência $\mathcal{E}$=[*alti* $\geq 10^4$] pelo grafo actualizado.

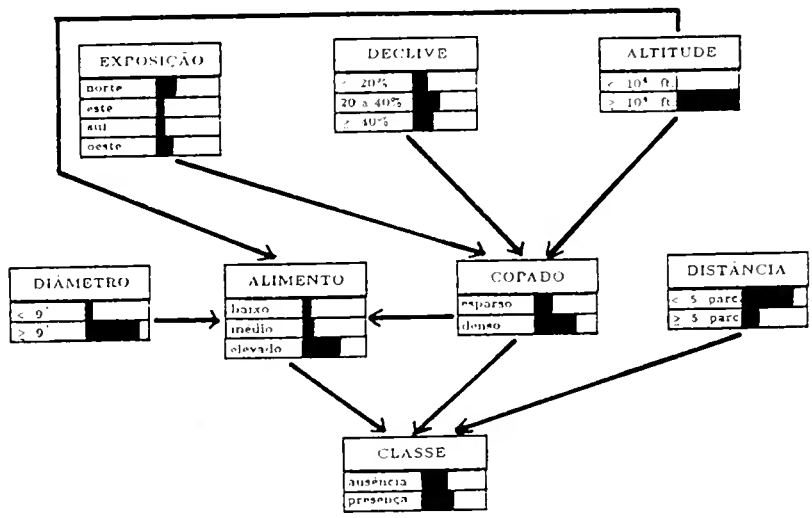


Figura 7.6 - Grafo bayesiano: propagação de $\mathcal{E}$=[*alti* $\geq 10^4$]

Os valores obtidos para $p(v_i|\mathcal{E},\mathcal{G})$, são condiderados aceitáveis. A modificação introduzida no modelo gráfico é mantida e o processo prossegue com uma nova ramificação da árvore de classificação.

É escolhida a variável *dist* que discrimina perfeitamente os objectos das duas classes na amostra. A evidência corrente passa a ser $\mathcal{E}$=[*alti* $\geq 10^4$] $\wedge$ [*dist* $\geq$ 5]. Ao nó criado está associado apenas um objecto pertencente à classe *ausência* e, portanto, a subamostra não é considerada representativa. $\mathcal{E}$ é propagada pelo grafo. A distribuição obtida para o nodo classe é $P(clas$ =*presença*$)$ = 0.7 e $P(clas$ =*ausência*$)$ = 0.3. O nó é afectado à classe que tem maior probabilidade.

A nova evidência para a ser então $\mathcal{E}$=[*alti* $\geq 10^4$] $\wedge$ [*dist* < 5]. A subamostra contém 8 objectos, todos da classe *presença*. A distribuição marginal obtida pela propagação de $\mathcal{E}$ pelo grafo é também favorável a essa classe, embora não de uma forma nítida ($P(clas$ =*presença*$)$ = 0.57 e $P(clas$ =*ausência*$)$ = 0.43). A relação entre as variáveis *dist* e *clas* não está bem esclarecido. O perito supõe que o aumento da distância às clareiras poderia, eventualmente, favorecer a presença de esquilos. Essa inconsistência (do tipo $dm \bowtie cp$) deve ter origem numa falha do modelo gráfico e terá que ser tomada em consideração no decorrer do processo de classificação.

Para que esta descrição do processo não se torne demasiado fastidiosa irão ser comentados apenas alguns passos (como foi referido, uma listagem integral encontra-se em anexo).

Ao analisar o nodo [*alti* $< 10^4$] a estimação, baseada na amostra, dos valores de

$P(v_i|\mathcal{E})$ para uma subamostra de 16 objectos produziu as distribuições de frequência apresentadas na figura 7.7.

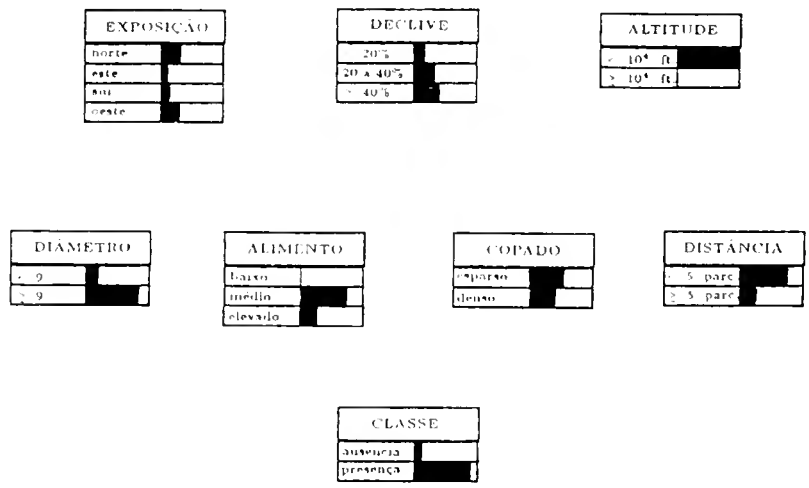


Figura 7.7 - Distribuições de frequências das variáveis na subamostra do nó $[alti < 10^4]$

Por achar que o grafo bayeiano deve ser modificado (pelas razões apresentadas aquando da introdução da aresta $alti \rightarrow alim$), o perito decide adicionar ao grafo a aresta $alti \rightarrow diam$. Após essa alteração os valores de $P(v_i|\mathcal{E},\mathcal{G})$ passam a ser os que constam na figura 7.8.

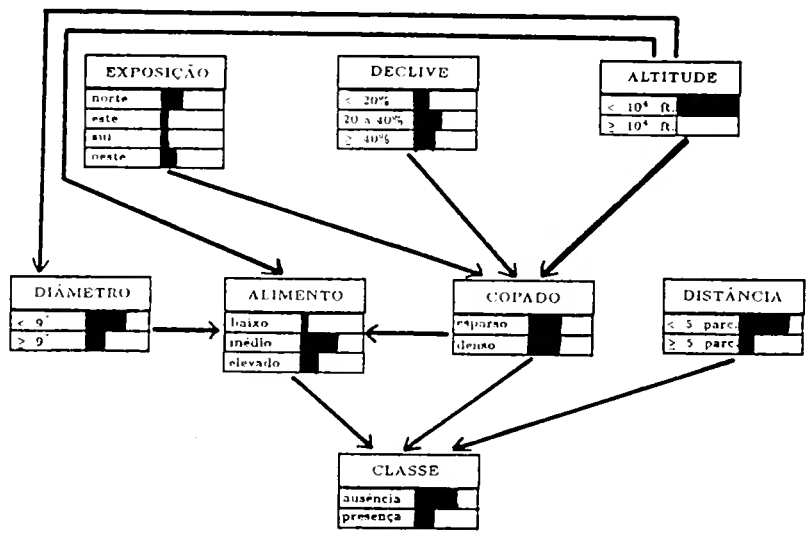


Figura 7.8 - Grafo bayesiano: propagação de $\mathcal{E}$=$[alti < 10^4]$

Como se pode verificar essa alteração conduz a estimativas (dadas pela subamostra e pelo grafo) diferentes no que respeita a distribuição de *diam*.

O processo de classificação prossegue até ser construida uma primeira árvore de classificação que é constituída pelas regras:

$[alti \geq 10^4\text{ft.}] \wedge [dist \geq 5parc.] \Rightarrow clas = ausência;$

$[alti \geq 10^4\text{ft.}] \wedge [dist < 5parc.] \Rightarrow clas = presença;$

$[alti < 10^4\text{ft.}] \wedge [alim = elevado] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = médio] \Rightarrow clas = ausência;$

$[alti < 10^4\text{ft.}] \wedge [alim = baixo] \Rightarrow clas = ausência.$

O processo de classificação continua com uma nova execução do módulo **iteração**. A árvore de classificação é construída desde o início mas o grafo bayesiano é o que resultou da iteração anterior do método.

As estimativas das distribuições marginais para $\mathcal{E}=[\ ]$ considerando o grafo bayesiano resultante da iteração anterior são dadas nas figuras 7.1 (as distribuições estimadas pela amostra são idênticas às da primeira árvore construida) e 7.9.

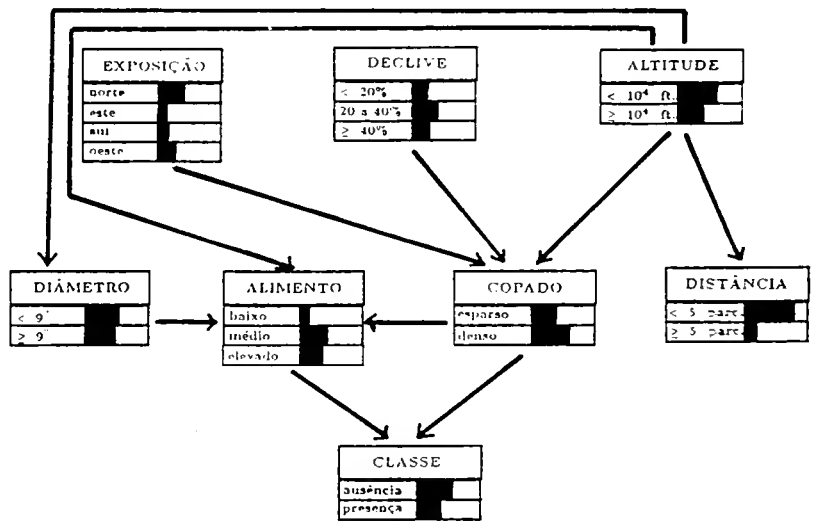


Figura 7.9 - Grafo bayesiano: propagação de $\mathcal{E}=[\ ]$

Verifica-se que há inconsistência no que respeita à variável *diam*. A aresta $alti \rightarrow diam$ que tinha sido anteriormente colocada é retirada. Note-se que a aresta $dist \rightarrow clas$ foi entretanto retirada, o que implica que $dist \perp\!\!\!\perp clas|alti$.

Na segunda iteração a escolha da variável responsável pela primeira ramificação recai sobre *alim* pois o valor da função informação em *alim* é apenas ligeiramente superior ao valor em *alti*.

O processo prossegue e, num determinado passo, $\mathcal{E}=[alim = elevado] \wedge [expo =$

$sul] \wedge [alti < 10^4]$, e a subamostra é constituída por dois objectos da classe 0. Pelo grafo bayesiano são estimadas as distribuições apresentadas na figura 7.10.

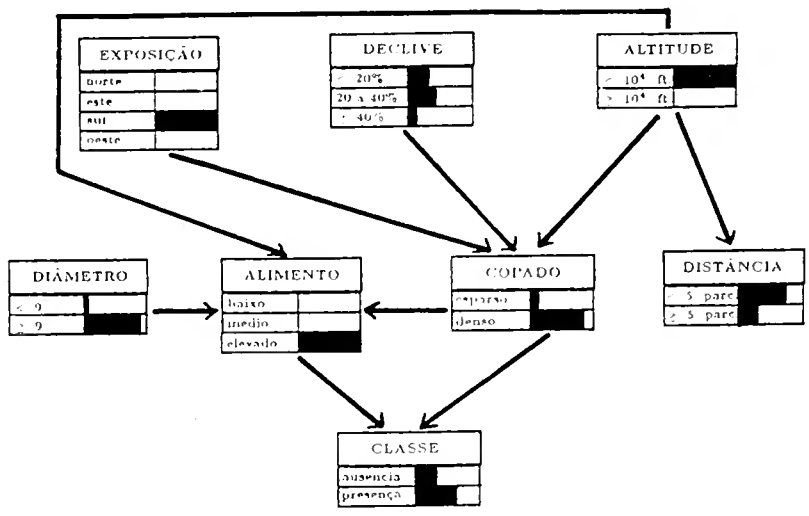


Figura 7.10 - Grafo bayesiano: propagação de
$\mathcal{E}$=$[alim = elevado] \wedge [expo = sul] \wedge [alti < 10^4]$

Embora o perito ponha em causa os valores obtidos, porque o seu conhecimento do problema indica que as condições estabelecidas em $\mathcal{E}$não são claramente favoráveis à presença de esquilos, decide afectar o nodo à classe *presença*. A amostra é considerada não representativa. O perito poderia ter decidido ramificar de novo o nó ou alterar o grafo bayesiano.

A última evidência considerada para a construção da árvore da segunda iteração é $\mathcal{E}$=[alim=baixo]. A subamostra correspondente é vazia e, consequentemente, a decisão a tomar dever-se-á apoiar no resultado da propagação de $\mathcal{E}$ pelo

grafo que é dada na figura 7.11.

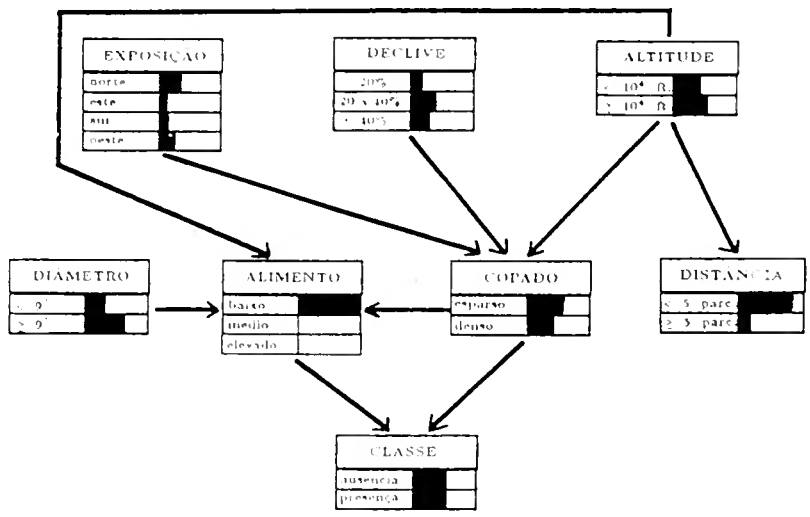


Figura 7.11 - Grafo bayesiano: propagação de $\mathcal{E}$=[*alim* = *baixo*]

Como se verifica, não é possível tomar uma decisão. A distribuição marginal obtida para *clas* é inconsistente com o conhecimento do perito. Este considera que as condições são claramente desfavoráveis à presença de esquilos quando o nível alimentar é baixo. Por essa razão decide afectar o nodo corrente à classe ausência. Esta inconsistência justificaria uma nova iteração.

A árvore final é constituída pelas regras:

[*alim* = *elevado*] $\wedge$ [*expo* = *norte*] $\Rightarrow$ *clas* =*presença*;

[*alim* = *elevado*] $\wedge$ [*expo* = *este*] $\wedge$ [*alti* $\geq 10^4$ft.] $\Rightarrow$ *clas* =*presença*;

[*alim* = *elevado*] $\wedge$ [*expo* = *este*] $\wedge$ [*alti* $< 10^4$ft.] $\Rightarrow$ *clas* =*presença*;

[*alim* = *elevado*] $\wedge$ [*expo* = *sul*] $\Rightarrow$ *clas* =*presença*;

[*alim* = *elevado*] $\wedge$ [*expo* = *oeste*] $\Rightarrow$ *clas* =*presença*;

[*alim* = *médio*] $\Rightarrow$ *clas* =*ausência*;

[*alim* = *baixo*] $\Rightarrow$ *clas* =*ausência*.

Como se pode verificar, esta árvore de classificação é equivalente a uma árvore mais simples com apenas três regiões de decisão pois todos os nós terminais que resultam da ramificação de $\mathcal{E}$=[*alim* = *elevado*] são afectados à classe *presença*.

7.3. Resultados

Os resultados obtidos irão ser analisados sob duas perspectivas. A primeira está ligada ao primeiro objectivo do método proposto que consiste em classificar o mais correctamente possível os objectos de classificação inicialmente desconhecida. A segunda perspectiva é a da integração do conhecimento elementar e suplementar e a captação desse conhecimento numa base de conhecimento sobre o domínio dos objectos.

A avaliação do possível interesse do método MCIC será realizada segundo essas duas perspectivas. O MCIC distingue-se dos métodos que lhe servem de comparação por ter objectivos adicionais.

A comparação dos métodos deve ser feita distinguindo, igualmente, as seguintes situações. Por um lado, três casos podem ocorrer na sequência dessa comparação: *MCIC* pode ter um desempenho inferior, igual ou superior relativamente à capacidade de classificar objectos de classe desconhecida. Por outro lado, a entidade interessada na resolução do problema pode ter duas posições: estar apenas interessada numa regra que classifique correctamente os objectos, ou (adicionalmente ao objectivo anterior) pretender ter um modelo mais completo que permita fazer uma análise mais diversificada do problema e/ou pretender ter uma regra que explique sob a forma de conceitos facilmente compreensíveis a regra de classificação e/ou pretender captar o conhecimento de um determinado perito sobre o domínio do problema.

Na primeira destas duas últimas situações, MCIC só se justifica se tiver um desempenho superior na classificação dos objectos de classe desconhecida. Na segunda, MCIC pode considerar-se útil, mesmo que tenha um desempenho inferior. Neste caso, seria preferível utilizar outro método para obter a regra de classificação.

7.3.1. Classificação dos objectos de classe desconhecida

As taxas que serão apresentadas para quantificar a eficiência da regra são, formalmente, taxas aparentes pois são calculadas sobre toda a população (incluindo a amostra). No entanto, fornecem uma medida muito semelhante à taxa real porque as amostras consideradas são de pequena dimensão relativamente ao efectivo da população.

São apresentadas na tabela 7.3 as percentagens de objectos bem classificados (afectados, pela regra de classificação, à classe a que realmente pertencem), mal classificados e não classificados (esta percentagem é de considerar quando a regra define

regiões de decisão não afectadas a nenhuma das classes consideradas à partida). Os métodos de classificação comparados são: 1) a regressão logística com definição de um *ponto de corte* entre as classes; 2) o método ID3; 3) o mesmo método com alteração do critério de afectação dos nodos *vazios*, ou seja, nós *vazios* afectados à classe mais provável segundo a distribuição obtida através do grafo bayesiano; 4) MCIC.

Na mesma tabela podem ser observadas medidas de complexidade para todos os casos. A complexidade é medida pelo número de regiões de decisão que a regra define. No caso das regras com forma de árvore a complexidade é dada pelo número de nós terminais.

métodos de classificação	dimensão da amostra	percentagem de objectos			complexidade
		bem classif.	não classif.	mal classif.	
regressão logística	50	72.6	-	27.4	2
	25	71.1	-	28.9	2
ID3	50	73.0	3.9	23.1	21
	25	66.9	11.0	22.1	10
ID3 "misto"	50	75.6	-	24.4	21
	25	76.6	-	23.4	10
MCIC	50	74.3	-	25.7	9
	25	74.7	-	25.3	3

Tabela 7.3 - Comparação dos métodos de classificação

Os valores da tabela 7.3 indicam, para o problema considerado, que: 1) os desempenhos dos métodos baseados em árvores não são inferiores aos do método paramétrico convencional; 2) o método "misto", que resulta de uma complementação da informação da amostra pela informação incorporada no final do processo de classificação no grafo bayesiano, teve bom comportamento; 3) o método proposto reduz significativamente a complexidade da regra de classificação originada pelas abordagens baseadas em árvores mantendo um desempenho comparável. Esta última diferença deve-se ao facto das regras para a tomada de decisão nos nós da árvore serem mais flexíveis do que as do método ID3, isto é, conduzirem a uma decisão de não ramificação de nós (e afectação) para nós que seriam ramificados por ID3.

Estes resultados devem ser analisados cautelosamente pois baseiam-se numa única experiência para cada caso. Seria interessante fazer uma comparação semelhante para médias das taxas de erro obtidas com diversas amostras aleatórias.

7.3.2. Caracterização do domínio dos objectos

Pelas descrições dos diversos métodos foi possível observar a forma como a regra de classificação caracteriza o domínio dos objectos e não parece ser polémico afirmar que os métodos baseados em árvores originam uma regra muito mais *explicativa* do que o método estatístico clássico. O método proposto, para além de se enquadrar nesse último caso, produz uma descrição muito mais pormenorizada do domínio dos objectos.

No que diz respeito à captação de conhecimento pericial, a execução do método MCIC coloca, em cada passo, ao perito novos problemas de formulação relativamente simples e leva-o, constantemente, a desenvolver raciocínios e tomar decisões. Pode-se questionar a *qualidade* do conhecimento armazenado. O método permite considerar informação factual, e portanto objectiva, e informação mais ou menos subjectiva proveniente do perito. A estratégia do método é dar preferência à primeira quando esta é representativa e fornecer um suporte para a estruturação da segunda em caso contrário. O método proposto oferece um suporte para estruturação do conhecimento pericial e torna-o comparável com outras formas de conhecimento, ou seja, torna-o *manipulável*.

7.3.3. Conclusões

Recorde-se o que foi dito no início de 7.3. Em relação ao exemplo apresentado pode-se tirar duas conclusões.

A primeira conclusão é que o método não deve ser preterido em relação aos outros métodos aplicados no que respeita ao objectivo de construir uma regra de classificação que classifique bem os objectos. Isso significa que a utilização de MCIC é vantajosa no caso de se pretender, para além desse objectivo, obter uma caracterização mais completa do domínio do problema. Existem três ordens de razões para essa afirmação: 1) MCIC origina uma regra de classificação mais explicativa do que a regra obtida pela regressão logística; 2) MCIC origina uma regra mais compreensível do que a regra obtida através de ID3 (devido à sua menor complexidade); e 3) MCIC fornece uma representação complementar do domínio do problema, o grafo bayesiano (cujo modelo gráfico tem uma interpretação muito simples).

A segunda conclusão é que, no caso de se pretender apenas uma regra de classificação eficiente, é preferível utilizar um método não interactivo e mais expedito. Na verdade, não se verificou, no exemplo tratado, que MCIC permitisse, apesar da incorporação de conhecimento pericial, construir uma regra de classificação mais

eficiente que os métodos de indução a partir da amostra.

Analisem-se agora as exigências computacionais dos diversos métodos para a aplicação realizada. Para a amostra de dimensão 50, o método que teve menores exigências foi ID3 (a execução exigiu aproximadamente 1 minuto de cpu, num computador VAX 4200). O método MCIC e a regressão logística tiveram exigências elevadas, tendo em conta a dimensão dos dados de entrada. A execução de qualquer um dos dois exigiu aproximadamente 10 minutos de cpu, no mesmo computador. No caso de MCIC isso pode ser explicado pela baixa eficiência do dialecto e versão da linguagem PROLOG utilizado. No caso da regressão logística a tempo de cálculo deve-se ao procedimento de selecção de variáveis e ao facto de, por transformação das variáveis discretas, terem sido consideradas quatorze variáveis (quatro contínuas e dez binárias), em vez das sete variáveis consideradas nos outros métodos.

8 . Conclusões e considerações finais

Os problemas de classificação supervisada são muito frequentes. Muitos problemas de modelação, simulação, previsão, e planeamento de tarefas podem ser formulados como problemas de classificação. Em todas as ciências aplicadas os métodos de classificação são ferramentas correntemente utilizadas.

Nas áreas mais diversas, como as ciências agrárias (como o domínio do exemplo tratado), a medicina (de onde provêm diversos exemplos referidos na bibliografia), as ciências sociais e a economia, existe um confronto com o *desconhecido*, com uma grande diversidade de objectos em estudo, pressões de ordem social, económica, ou outras, que conduz à necessidade de encontrar classes e conceitos representativos, encontrar categorias e enquadrar a realidade em campos bem definidos. Desse modo é obtida uma primeira aproximação para a compreensão dos fenómenos em questão.

Essa aproximação pode ser restritiva em relação à realidade se os modelos utilizados para a representação do conhecimento não forem adequados à complexidade dessa realidade.

Por um lado, existem modelos projectados para suportar conhecimento muito estruturado como sejam os modelos analíticos, determinísticos, cujo nível de representação da realidade consiste em entidades complexas. Por outro lado, existem modelos que representam conceitos simples, e que podem servir de suporte ao conhecimento geral e pouco especializado sobre um determinada área de saber.

Evidentemente, entre esses extremos existem métodos específicos para determinado nível de representação da realidade ou para um determinado problema restrito.

O que se tentou fazer, no trabalho apresentado, foi integrar no mesmo método dois modelos de representação, que podiam representar a mesma realidade segundo duas perspectivas distintas. Os dois modelos foram, como foi amplamente referido, uma árvore de classificação e um grafo bayesiano. Por forma a conseguir realizar a integração pretendida, foram utilizadas métodos e técnicas de disciplinas como a Inteligência Artificial, a Estatística e a Investigação Operacional.

Os dois modelos escolhidos situam-se num nível intermédio de especialização. A árvore de classificação tem um nível mais simples de representação pois suporta uma hierarquia de conceitos simples. O grafo bayesiano, embora tenha uma estrutura globalmente complexa, é composto por módulos individuais de pequena dimensão e complexidade.

Como foi referido, o mesmo problema (por exemplo, o problema analisado em 7) é descrito segundo perspectivas distintas: a árvore representa os objectos como objectos simbólicos, pertencentes a uma hierarquia; e o grafo representa os objectos como relações de dependência entre as variáveis que os descrevem.

O método MCIC propõe uma forma de integração a dois níveis das duas estruturas. A integração passa pela criação de formas de comunicação, ou seja, formas de troca de informação entre as estruturas. Esse objectivo foi atingido, por um lado, através da noção de independência condicional entre variáveis e classes e, por outro, através de um modelo de efeitos entre variáveis, um modelo "propagativo" no caso do grafo correspondente a um modelo "aditivo" (resultante de uma especialização de um conceito por acção de uma variável) no caso da árvore de classificação.

A relação teóricamente estabelecida suporta a construção de um método operacional para a obtenção da regra de classificação para o problema considerado.

Para além do aspecto da integração de informação, e devido ao facto de o grafo bayesiano ser um modelo adequado á estruturação de conhecimento pericial, o método de classificação adquiriu uma nova função: servir de enquadramento para a incorporação do conhecimento pericial no modelo. Verificou-se, através do exemplo, que a execução do método mostrou que essa forma de incorporação do conhecimento do perito pode ser eficiente. O perito, ao longo do processo, é colocado em frente a problemas que, por serem colocados sob duas perspectivas distintas (segundo a árvore, pondo em foco os conceitos, e segundo o grafo, pondo em evidência as relações entre variáveis) levam-no a desenvolver raciocínios e a colocar e avaliar hipóteses que são captadas (e, eventualmente, revistas em novas situações).

Em contrapartida, o exemplo apresentado não foi muito conclusivo, e até um pouco desanimador, sobre as potencialidades do método MCIC para a construção da regra de classificação. Este facto pode dever-se ao problema particular considerado. O método deveria ser aplicado a outros problemas para se poder avaliar mais justamente o seu desempenho.

O trabalho desenvolvido para elaboração do método proposto, para além de um motivo para estudar algumas áreas da Matemática Aplicada, foi um desafio no sentido em que os objectivos colocados à partida e a gama de problemas de classificação considerados colocavam dificuldades na definição das técnicas e métodos que, por um lado não fossem restritivos, e por outro, fossem operacionais.

O método cumpriu parcialmente os objectivos fixados à partida. O método

proposto permite dar uma resposta a esses objectivos (construir uma regra de classificação e integrar e estruturar conhecimento pericial sobre o domínio dos objectos) e, pode dizer-se que os aspectos do conhecimento, simbolismo e raciocínio referidos na introdução foram contemplados.

O método impõe limitações sobre a diversidade de conhecimento que pode ser manipulado, particularmente, quando implica a discretização das variáveis envolvidas e tem um comportamento fraco num aspecto importante: é muito pouco automatizado pois, na sua forma actual, depende fortemente da interacção com o perito para a sua execução. O método proposto pode ser, nesse aspecto, melhorado. Pode ser definido, com maior precisão, o que se entende por inconsistência entre a estimativa das distribuições marginais obtida através da amostra e através do grafo. O método poderia então auxiliar melhor o perito fornecendo indicações precisas sobre os parâmetros a considerar e poderia tornar-se mais automatizado.

9 . Bibliografia

BREIMAN, L., FRIEDMAN, J.H., OLSHEN, R.A. and STONE, C.J. (1984). Classification and Regression Trees. Wadsworth and Brooks/Cole Advanced Books and Software, Pacific Grove, California.

COELHO, H. e COSTA, E. (1992). Elementos de Inteligncia Artificial. (a publicar).

DAWID, A.P. (1979). Conditional Independence in Statistical Theory. Journal of the Royal Statistical Society, Series B, 41, pp 1-31.

DIDAY, E. (1989). Introduction à l'analyse des données symboliques. Rapports de Recherche, INRIA, nº1074.

GEIGER, D., VERMA, T. and PEARL, J. (1990). Identifying independence in bayesian networks. Networks, vol. 20, pp 507-534.

GNANADESIKAN, R., BLASHFIELD, R.K., BREIMAN, L., DUNE, O.J., FRIEDMAN, J.H., KING-SUN FU, HARTIGAN, J.A., KETTENRING, J.R., LACHENBRUCH, P.A., OLSHEN, R.A. and ROHLF, F.J. (1989). Discriminant analysis and clustering. Statistical Science, vol. 4, nº1, pp. 34-69.

HAND, D.J. (1981). Discrimination and Classification. John Wiley and Sons, New York.

KOWALSKI, R. (1979). Logic for Problem Solving. Artificial intelligence series, Nils J. Nilsson (ed.), Elsevier Science Publishing Co., Inc., New York.

LAURITZEN, S.L., DAWID, A.P., LARSEN, B.N., LEIMER, H.-G. (1990) Independence properties of directed Markov fields. Networks, 20, 491-505.

MIDDELKOOP, H. , JANSSEN, L.L.F. (1991). Implementation of temporal relationships in knowledge based classification of satellite images. Photogrammetric Engeneering and Remote Sensing, vol. 57, nº7,pp 937-945.

MILTON, J.S. e ARNOLD, J.C. (1989) Probability and Statistics in the Engineering and Computing Sciences. McGraw-Hill.

MOLLER-JENSEN, L. (1990). Knowledge-based classification of an urban area using texture and context information in landsat-TM imagery. Photogrammetric Engeneering and Remote Sensing, vol. 56, nº6, pp 899-904.

MOORE, D.M., LEES, B.G. , DAVEY, S.M. (1991). PROFILE - A new method for predicting vegetation distributions using decision tree anal-

ysis in a geographic information system. Environmental Management, vol.15, nº1, pp 59-71.

MURTEIRA, B.J.F. (1979). Probabilidades e Estatística, vol. I. McGraw-Hill de Portugal Lda.

PEARL, J. (1986). Fusion, propagation and structuring in belief networks. Artificial Intelligence, 29, 241-88.

PEARL, J. (1987). Evidencial reasoning using stochastic simulation. Artificial Intelligence, 32, 245-57.

PEREIRA, J.M.C. (1989). A spatial approach to statistical habitat suitability modelling: the Mt. Graham red squirrel case study. Unpublished Ph. D. Dissertation, School of Renewable Natural Resources, University of Arizona, Tucson, Arizona. 151 pp.

QIAN, J., EHRICH, R.W. , CAMPBELL, J.B. (1990). DNESYS - An expert system for automatic extraction of drainage networks from digital elevation data IEEE Transactions on Geosciences and Remote Sensing, vol. 28, nº1, pp 29-45.

QUINLAN, J.R. (1986). Induction of decision trees. Machine Learning 1, pp 81-106.

RICH, E. (1983). Artificial Intelligence. McGraw-Hill, Inc., Singapore.

SHAFER, G. (1987). Probability judgment in artificial intelligence and expert systems. Statistical Science, vol. 2, nº1, 1987, pp 3-44.

SKIDMORE, A.K. (1989). An expert system classifies eucalipt forest types using thematic mapper data and a digital terrain model. Photogrammetric Engeneering and Remote Sensing, vol. 55, nº10, pp 1449-1464.

SPIEGHELHALTER, D.J., DAWID, A.P., LAURITZEN, S.L., COWELL,R.G. (1992). Bayesian analysis in expert systems. Research report 92-6, MRC Biostatistics Unit., Cambridge.

SPIEGHELHALTER, D.J., LAURITZEN, S.L. (1990). Sequential updating of conditional probabilities on directed graphical structures. Networks, 20, 579-605.

SRINIVASAN, A., RICHARDS, J.A. (1990). Knowledge-based techniques for multi-source classification. International Journal of Remote Sensing, vol. 11, nº3, pp 505-525.

STOCKWELL, D.R.B., DAVEY, S.M., DAVIS, J.R., NOBLE, I.R. (1990). Using induction of decision trees to predict greater glider density. AI applications, vol.4, nº4, pp 33-43.

THOMPSON, B. , THOMPSON, W. (1986). Finding rules in data, an algo-

rithm for extracting knowledge from data. Byte, November, pp 149-158.

TOMASSONE, R., DANZART, M., DAUDIN, J.J. and MASSON, J.P. (1988). Discrimination et Classement. Masson, Paris.

Outras referências bibliográficas consultadas:

CHARNIAK, E. e MCDERMOTT, D. (1985). Introduction to Artificial Intelligence. Addison-Wesley Publishing Co., U.S.A.

CLOCKSIN, W.F. and MELLISH, C.S. (1987). Programming in Prolog, 3rd ed. Springer-Verlag, Berlin.

LINDLEY, D.V. (1987). The probality approach to the treatment of uncertainty in artificial intelligence and expert systems. Statistical Science, vol. 2, n 1, 1987, pp 3-44.

MISHKOFF, H.C. (1985). Understanding Artificial Intelligence. Sams understanding series, Howard W.Sams and Co., U.S.A.

PAPADIMITRIOU, C.H. e STEIGLITZ, K. (1982). Combinatorial Optimization: Algorithms and Complexity. Prentice-Hall Inc., New Jersey, U.S.A.

SPIEGELHALTER, D.J. (1987). Probabilistic expert systems in medicine: practical issues in handling uncertainty. Statistical Science, vol. 2, n 1, 1987, pp 3-44.

THORNTON, C.J. (1992). Techniques in Computational Learning. An introduction. Chapman and Hall (publ.), Great Britain.

Anexos

anexo A. - *Output* de uma execução do programa PCIC para uma amostra de 25 objectos. As variáveis e as suas modalidades são representadas de uma forma diferente da do capítulo 7. As correspondências são as seguintes:

Anexo A	capítulo 7.
ac_elem(elev,cont,[0,10000])	alti $< 10^4$ ft.
ac_elem(elev,cont,[10000,20000])	alti $\geq 10^4$ ft.
ac_elem(slope,cont,[0,20])	decl $< 20\%$
ac_elem(slope,cont,[20,40])	decl $20a40\%$
ac_elem(slope,cont,[40,181])	decl $\geq 40\%$
ac_elem(dist,cont,[0,5])	dist < 5
ac_elem(dist,cont,[5,100])	dist ≥ 5
ac_elem(food,disc,[1,2])	alim=baixo
ac_elem(food,disc,[3])	alim=médio
ac_elem(food,disc,[4])	alim=elevado
ac_elem(asp,disc,[1])	expo=norte
ac_elem(asp,disc,[2])	expo=este
ac_elem(asp,disc,[3])	expo=sul
ac_elem(asp,disc,[4])	expo=oeste
ac_elem(copado,disc,[1,2])	copa=esparso
ac_elem(copado,disc,[3,4])	copa=denso
ac_elem(dbh,disc,[1,2,3])	diam $< 9''$
ac_elem(dbh,disc,[4])	diam $\geq 9''$